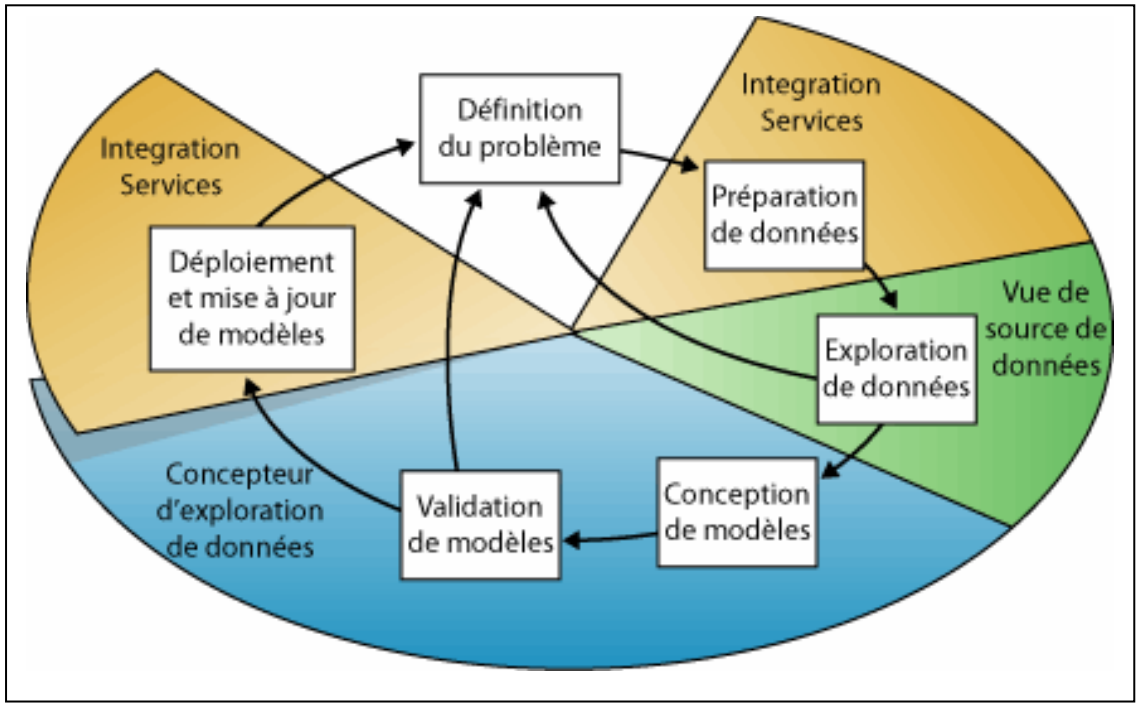
Hes·so VALAIS WALLIS
Haute Ecole Spécialisée de Suisse occidentale
Fachhochschule Westschweiz
University of Applied Sciences Western Switzerland

Travail de diplôme 2007

Filière Informatique de gestion

Facturation par APDRG : prédiction des recettes des cas non codés

PrediRec



Etudiant : Mathieu Giotta

Professeur : Henning Mueller

Projet SIMAV

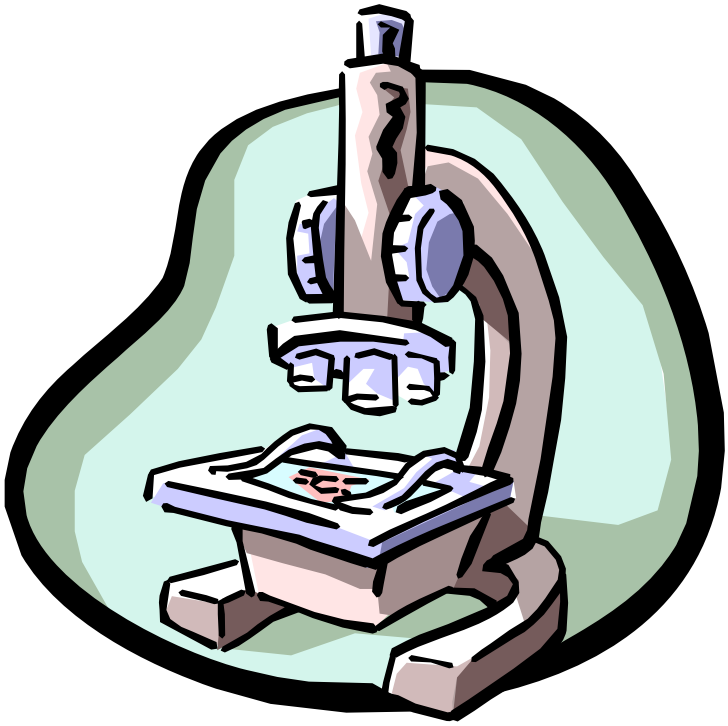
Rapport Final

Facturation par APDRG : prédiction des recettes des cas non codés

Institut Central des Hôpitaux Valaisans
Av. Grand Champsec 86
Case postale 736
1951 Sion
Suisse

Auteur	: Mathieu Giotta	**Date de création**	: 25.11.2007
Fichier	: PrediRec - 002 Rapport de projet	**No de version**	: V finale
Etat	:	**Dernière révision**	:
Distribution	:	**Date de distribution**	:
Publication	:		

Table des matières



1 Introduction

1.1 Présentation du mandant

Le SIMAV est l'acronyme de Service d'Informatique Médicale et Administrative Valaisan.
Ce service est intégré au Réseau Santé Valais (RSV) et doit fournir l'appui informatique nécessaire au bon fonctionnement des hôpitaux valaisans. Ce service est dirigé par le Dr. Gnaegi.

Le SIMAV effectue les développements, la gestion et la maintenance de l'infrastructure informatique du RSV. Parmi ses nombreuses tâches, nous pouvons citer :

- maintenance réseau ;
- maintenance matérielle ;
- support aux utilisateurs (back office) ;
- développement de petites solutions.

C'est dans cette dernière activité que s'inscrit ce projet.

1.2 Contexte du projet

Depuis 2005, tous les séjours hospitaliers en soins somatiques aigus (médecine, chirurgie, gynécologie-obstétrique, pédiatrie, etc.) sont facturés par le RSV sous forme de forfaits liés à la pathologie (APDRG pour All Patients Diagnosis Related Groups). Cette méthodologie de facturation remplace celle des forfaits journaliers qui était utilisée auparavant. La génération de ces forfaits APDRGs implique le codage préalable des diagnostics et des interventions documentés dans le dossier médical du patient. Or, lors du bouclement comptable des hôpitaux valaisans, tous les patients sortis durant l'exercice terminé ne sont pas forcements codés, et ne peuvent donc pas être facturés.

Il est alors nécessaire de provisionner les recettes, qui seront perçues après bouclement, afin de les intégrer à l'exercice comptable en cours. Cependant, comme ces recettes dépendent de la pathologie, elles ne sont à priori pas connues.

1.3 But du développement

Le but de ce projet est de déterminer un modèle d'analyse afin d'estimer au mieux les recettes liées aux cas non codés à partir des informations disponibles dans les systèmes opérationnels.
Pour parvenir à provisionner les recettes de ces cas, il a été décidé de mettre en place une solution de Data Mining, que nous appellerons « PrediRec » dans le reste du document.

1.4 Description de l'existant

Actuellement, les hôpitaux calculent leurs provisions pour le secteur somatique aigu de manière empirique sur la base d'un montant moyen forfaitaire.
Cette solution ne donne pas entière satisfaction, car elle ne tient pas compte de la lourdeur des cas et, de plus, cette méthode est assez fastidieuse.

2 Description générale des buts de PrediRec

PrediRec (Concaténation des termes Prediction et Recette) est une solution de Data Mining permettant de simuler les recettes liées aux cas non codés des hôpitaux valaisans et pour lesquels les comptables doivent effectuer des prévisions lors des bouclements comptables trimestriels ou annuels.

Etant donné que les différents utilisateurs sont des comptables et des facturistes, nous avons décidé de créer une interface utilisateur simplifiée qui permette une utilisation intuitive de l'outil sans formation préalable. Le fait d'utiliser un site Internet est un avantage car il nécessite aucune installation sur les postes clients et il est disponible depuis n'importe quel ordinateur du RSV.

L'application PrediRec est déposée dans un portail Intranet du SIMAV qui est déjà connu des utilisateurs finaux de PrediRec. Cet Intranet est celui qui héberge le Data Warehouse et, comme ce projet est développé au sein du groupe en charge du Data Warehouse, la maintenance s'en trouve ainsi simplifiée.

Comme le point fort de ce travail de diplôme est la maîtrise des outils et concepts du Data Mining et de son application/intégration en entreprise, les fonctionnalités et l'ergonomie de l'interface utilisateur ont été réduites à un minimum, mais sont suffisantes pour que l'application PrediRec soit utilisable.

2.1 Vue d'ensemble des fonctionnalités

Le site Internet de PrediRec est en mesure de proposer les fonctionnalités suivantes :

Choix de la variable à estimer
L'utilisateur doit sélectionner quelle est la variable qu'il désire estimer :

- soit le « Cost-Weight (CW) pondéré » ;
- soit le « Total Facture ».

Mise à jour des modèles d'exploration de données
Par cette opération, l'utilisateur effectue la phase d'apprentissage nécessaire à l'application de Data Mining.
Chaque utilisateur peut créer son propre modèle, c'est-à-dire qu'un utilisateur « x » peut mettre à jour son modèle de données pendant que l'utilisateur « y » travaille sur le sien.
De cette manière, un utilisateur peut évaluer ses cas non codés sans risquer d'être déstabilisé par une mise à jour impromptue du modèle.

Choix des cas non codés
L'utilisateur final peut sélectionner les cas non codés qu'il désire estimer. PrediRec se connecte à la base de données du Data Warehouse (DW) pour en extraire les informations relatives aux cas choisis.

Estimation des cas choisis
La principale fonctionnalité de PrediRec repose sur l'estimation des cas non codés choisis. Par cette action, l'utilisateur soumet ces cas au moteur de Data Mining qui les lui retourne, complétés par son évaluation.

Exportation des cas simulés dans MS Excel
A la suite d'une simulation de cas non codés, l'utilisateur peut exporter les résultats dans MS Excel.

Simulation d'un cas fictif
L'utilisateur de PrediRec a la possibilité de créer un cas fictif dans le but, par exemple, d'effectuer un devis avant une hospitalisation.

Choix de la langue de l'application Web
Etant donné que cette application peut être utilisée par des francophones ainsi que par germanophones, l'utilisateur est libre de choisir dans quelle langue les textes sont affichés.

2.2 Plan

Il a été décidé de rendre le rapport final du travail de diplôme pour le 21 décembre 2007

Tâches	Échéance	Personne responsable	Effective
Validation du cahier des charges	02/11/2007	HMU, AGN	02/11/2007
Validation de l'analyse	13/11/2007	AGN, TWE	20/11/2007
Choix du modèle de Data Mining	04/12/2007	AGN, TWE	11/12/2007
Fin du travail de diplôme	21/12/2007	MGI	21/12/2007
Défense du travail de diplôme	08/01/2008	HMU, MGI	

2.3 Environnement technique

Le moteur de Data Mining est installé sur la même machine que celui du Data Warehouse. Le serveur en question possède un processeur AMD Opteron cadencé à 2.83 GHz 64 bit, greffé de 16 GB de RAM.

Le système de gestion de base de données relationnelle (SGBDR) du Data Warehouse est une base de données SQL Server 2005 et celui de PrediRec également

Les pages Internet de PrediRec sont quant à elle placées dans le portail Web du Data Warehouse et qui se nomme Business Objects Infoview XI, communément appelé WebI. Il s'agit d'un site Internet fonctionnant sur un serveur sécurisé (SSL) IIS 5. Ce site Internet est hébergé sur un serveur possédant un processeur Intel Xeon 3.066 GHz greffé de 2.5 GB de RAM.

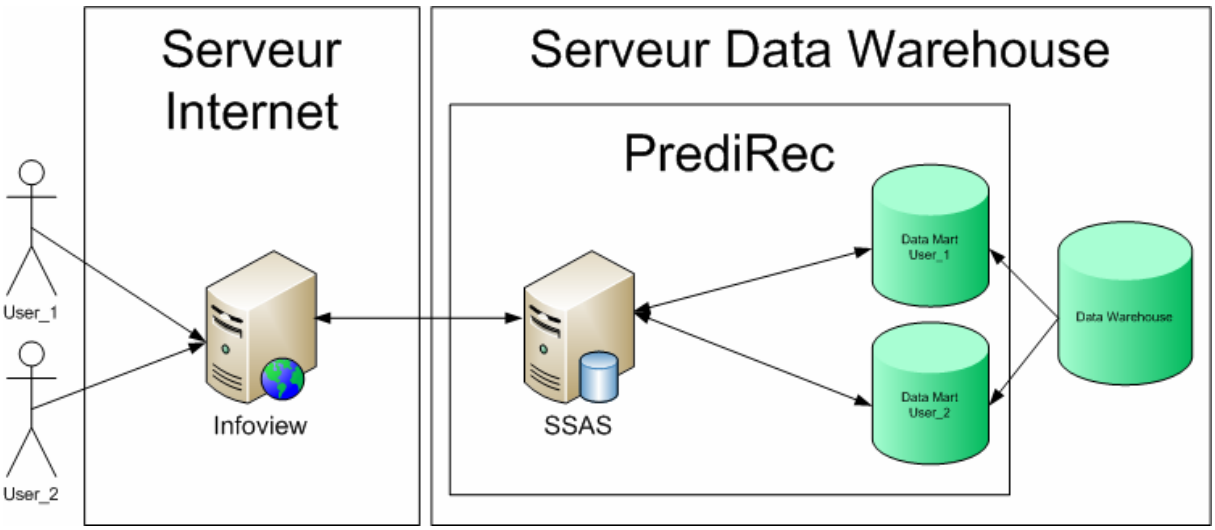


Figure 2-1 : Environnement technique

3 Présentation des techniques de Data Mining

3.1 Définition

Qu'est ce que le Data Mining ?
Le terme Data Mining regroupe des techniques d'analyse utilisant un logiciel pour établir des tendances ou des corrélations cachées parmi des masses de données, ou encore pour détecter des informations stratégiques ou découvrir de nouvelles connaissances, en s'appuyant sur des méthodes de traitement statistique.[1]

En d'autres termes, les techniques de Data Mining permettent d'extraire des règles, des associations ou des informations « cachées » à partir des données stockées en entreprise, données souvent inexploitées.

A partir des règles trouvées, nous pouvons, par exemple, définir des profils d'acheteurs, des tendances sur leurs achats, etc. Les informations ainsi révélées peuvent être appliquées, par exemple, lors d'un publipostage ciblé qui permettrait d'économiser de l'argent en évitant d'envoyer de la publicité à des acheteurs qui ne sont pas intéressés ou lors de l'attribution d'un emprunt bancaire estimé si l'emprunteur possède un profil à risque ou non.

Dans le cadre de ce projet, nous voulons définir si il est possible d'extraire à partir des systèmes sources des règles ou des corrélations sur les séjours des patients et les recettes liés à ceux-ci.

3.2 Méthodologies du Data Mining

Un projet de Data Mining doit, en général, se dérouler en six étapes distinctes :

- la définition du problème ;
- la préparation des données ;
- la création du modèle d'analyse ;
- l'exploration du modèle d'analyse ;
- la validation du modèle d'analyse ;
- le déploiement du modèle d'analyse.

Durant le projet, il peut être nécessaire de recommencer les étapes de création, d'exploration ainsi que la validation du modèle d'analyse, voir même recommencer la définition du problème.

[1] http://www.journaldunet.com/encyclopedie/definition/204/51/20/datamining.shtml

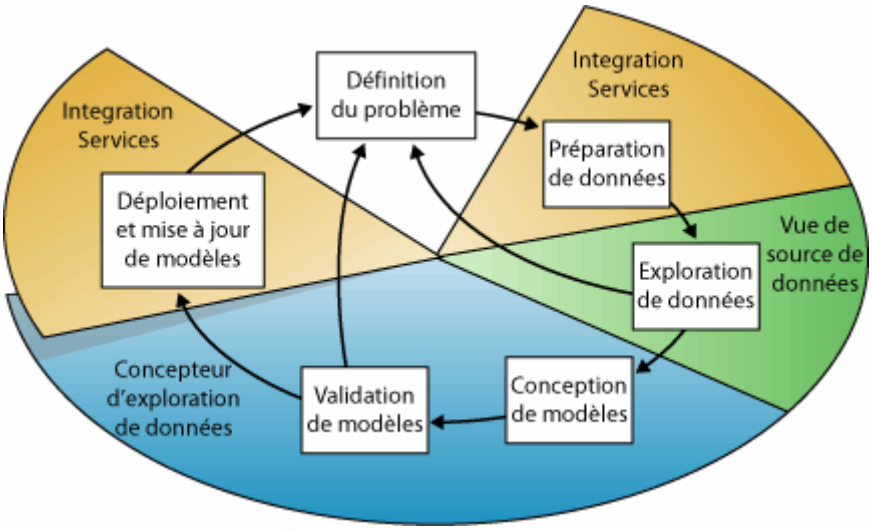


Figure 3-1 : Déroulement d'un projet de Data Mining

3.2.1 Définition du problème

La première étape consiste à définir le problème qui doit être résolu.
Par exemple :

- mise en place d'un publipostage ciblé ;
- création d'un « panier d'achat » ;
- activité sur d'un site web ;
- approximation des cas non codés lors d'un bouclement comptable (ce projet) ;
- etc.

Cette étape est en général sollicitée par les dirigeants d'une société.

3.2.2 Préparation des données

La seconde étape consiste à réunir dans une même base de données, les diverses sources disponibles de la société.

Dans les sociétés, il existe généralement plus d'une source de données qui peut contenir des informations pouvant être pertinentes. Avant de commencer un projet de Data Mining, il est donc nécessaire d'utiliser un outil d'ETL (Extraction, Transformation and Load) afin de réunir toutes ces données en un même point, de les nettoyer et de les normaliser.

Durant cette étape, la connaissance des données sources est primordiale.

Une fois les données collectées, nous les séparons en deux sets de données :

- le premier set correspond aux données avec lesquelles nous allons construire le modèle d'analyse. Nous nommons ce set : « données d'apprentissage » ;
- le deuxième set servira à évaluer le modèle d'analyse. Nous appelons ce set : « Données de test ».

Le set de données de test peut être aussi un échantillonnage des données d'apprentissage, mais nous conseillons de séparer dès le départ les deux jeux de données. De cette manière, nous sommes certains que, lors de l'évaluation du modèle, celui-ci ne soit pas faussé à cause de l'utilisation de cas particuliers et uniques qui ont servi lors de la phase d'apprentissage.

3.2.3 Création du modèle

La troisième étape consiste à créer un modèle d'analyse, c'est à dire que l'on demande au logiciel de Data Mining qu'il analyse et traite les données d'apprentissage et qu'il nous fournisse un modèle d'analyse.

Durant cette étape, il est nécessaire de spécifier quelle est l'utilité des variables fournies au logiciel de Data Mining.
Nous devons lui spécifier quelles sont :

- la clé des enregistrements (clé primaire – clé unique)
- les attributs d'entrées
- les variables prévisibles (selon le modèle d'analyse sollicité[i]).

Nous devons aussi spécifier quel(s) est (sont) le(s) algorithme(s) à utiliser.

i Certains modèles d'analyse sont utilisés uniquement dans le but de trouver des associations entre les variables.

3.2.4 Exploration du modèle

La quatrième étape consiste à explorer les analyses effectuées par l'outil de Data Mining.

Durant cette étape, l'outil de Data Mining permet de « naviguer » dans le modèle d'analyse. Dans SSAS (SQL Server Analysis Services), nous utilisons la « Visionneuse d'exploration de données ». Cet utilitaire nous permet d'afficher, selon l'algorithme utilisé, un réseau de dépendance qui nous permet d'observer comment le modèle a utilisé les variables que nous lui avons fournies.

3.2.5 Validation du modèle

La cinquième étape consiste à valider le modèle d'analyse proposé par le logiciel.

Pour effectuer cette étape, nous exécutons le modèle d'analyse proposé à l'étape précédente avec des données de test pour lesquelles nous connaissons déjà les résultats. Ensuite, nous comparons les résultats proposés par rapport aux résultats attendus. Et nous définissons si le modèle répond à nos attentes ou si nous devons redéfinir les attributs d'entrée (tâche « 3.2.3 Création du modèle »), voir même reformuler le problème (tache « 3.2.1 - Définition du problème »).

3.2.6 Déploiement et mise à jour du modèle

La sixième et dernière étape consiste à la mise en production du modèle d'analyse.

Lorsque le modèle est validé, il doit être déployé sur le serveur de production et, selon l'environnement, des rôles d'utilisateurs doivent être définis.

La mise à jour du modèle doit être effectuée dès l'introduction de nouvelles données dans le(s) système(s) source(s). Cette tache peut être confiée à un outil ETL.

3.3 Tâches du Data Mining

Les chapitres 3.3.1 à 3.3.5 sont repris du livre « Business Intelligence avec SQL Server 2005, Mise en œuvre d'un projet décisionnel »[1].

3.3.1 Classification

La classification consiste à examiner un objet afin de lui attribuer une classe. Pour le faire, l'outil de Data Mining se base sur des valeurs discrètes le composant (Ex. sexe, classe d'âge),

3.3.2 Régression

Une tâche de régression consiste à déterminer une relation entre des colonnes de type continu. L'outil de Data Mining génère une équation représentant au mieux la droite d'une série de données (Figure 3-2). L'équation contient le coefficient de corrélation ainsi que la covariance et affiche le R2. (Ex. y=0.958x – 754.23 : le coefficient de corrélation est 1.0982 et la covariance est de 227.6 pour un R2 de 0.6462)

Lorsque l'on effectue une régression, il est important d'analyser le R2 en résultant.
Le R2 ou coefficient de détermination mesure la qualité de l'ajustement des estimations de l'équation de régression. Il permet d'avoir une idée globale de l'ajustement du modèle. Il s'interprète comme la part de la variance de la variable Y expliquée par la régression, varie entre 0 et 1 et s'exprime souvent en pourcentage. En régression simple, un R2 proche de 1 est suffisant pour dire que l'ajustement est bon.

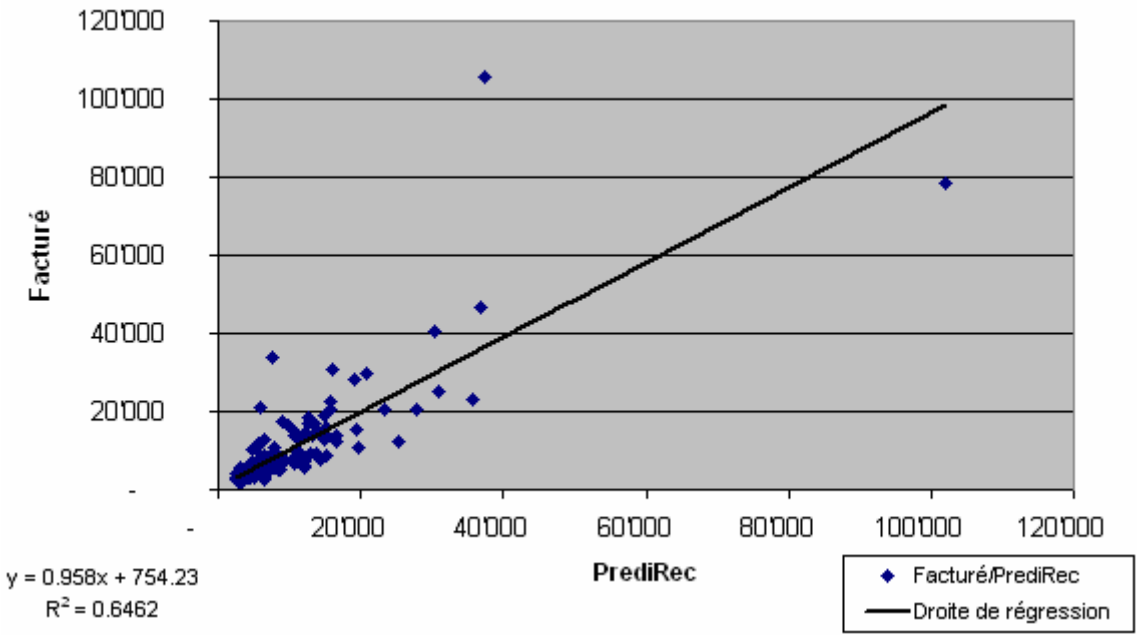


Figure 3-2 : Droite de régression

3.3.3 Segmentation – Clustering

Une tâche de segmentation consiste à créer des groupes d'objets homogènes depuis un set d'objets de premier abord non homogènes. Une tâche de segmentation est utilisée lors d'un processus de classification.

3.3.4 Association

Une tâche d'association examine les agissements d'un groupe d'individus pour déterminer des relations existantes entre eux.

Ex. Lors d'un achat sur un site de vente en ligne, le site Internet propose une liste d'articles qui ont été achetés lors d'achat d'autres personnes qui ont acquis le même objet.

3.3.5 Analyse de séquence

Une tâche d'analyse de séquence permet d'analyser une succession d'événements afin d'en prédire le futur.

Ex. On peut analyser l'enchaînement de pages Internet d'un site web et ainsi l'optimiser selon les habitudes des clients.

3.4 Technique du Data Mining

Les chapitres 3.4.1 à 3.4.5 sont repris du livre « Data Mining, Gestion de la relation client, Personnalisation de sites Web »[2].

3.4.1 Le raisonnement à base de cas

Les systèmes de raisonnement à base de cas (RBC) résolvent des problèmes par la comparaison d'exemples proches puisés dans un ensemble de cas préalablement stockés. Avec cette méthode de résolution, si une expérience passée et une nouvelle situation sont suffisamment « similaires », toutes les conclusions appliquées à l'expérience passée restent valides et peuvent être appliquées à la nouvelle situation.
Les RBC mettent en œuvre une procédure de recherche pour comparer les descriptifs du cas à traiter avec ceux des cas existants dans leur base interne. A ce titre, la capacité de résolution augmente au fil de l'arrivée de nouveaux exemples. Plus le nombre d'exemples est important, plus le RBC a de chances de retrouver un exemple proche, voir similaire.

3.4.2 Les arbres de décision

Un arbre de décision est un enchaînement hiérarchique de règles logiques construites automatiquement à partir d'une base d'exemples. Un exemple est constitué d'une liste d'attributs, dont la valeur détermine l'appartenance à une classe donnée. La construction de l'arbre de décision consiste à utiliser les attributs pour subdiviser progressivement l'ensemble d'exemples en sous-ensembles de plus en plus fins.

3.4.3 Les algorithmes génétiques

Les algorithmes génétiques ont été introduits par John Holland en 1975[2], avec la présentation d'une méthode d'optimisation inspirée de l'observation des capacités d'adaptation et d'évolution des espèces. Il a construit un système artificiel qui s'appuyait sur les principes de sélection de Darwin et sur les méthodes de combinaison des gènes de Mendel.
Les algorithmes génétiques renvoient aux principaux mécanismes de la sélection naturelle, c'est-à-dire essentiellement la sélection, la reproduction et la mutation. Les algorithmes génétiques décrivent l'évolution, au cours de générations successives, d'une population d'individus en réponse à son environnement. Ils sélectionnent les individus selon le principe de la survie du plus adapté.
Les algorithmes génétiques travaille sur une population de nombreuses solutions potentielles, toutes différentes. Le processus conduit à l'élimination des éléments les plus faible pour favoriser la conservation et la reproduction des individus les plus « performants » (les plus « justes »).

3.4.4 Les réseaux bayésiens

Les réseaux bayésiens sont une méthode classique utilisée pour associer une probabilité d'appariation d'un événement étant donné la connaissance de certaines autres probabilités. Un réseau bayésien est un modèle graphique qui encode les probabilités entre les variables les plus pertinentes.

[2] http://www.seas.upenn.edu/whatsnew/penderaward/jh.html

3.4.5 Les réseaux de neurones

Il nous faut d'abord distinguer les deux grandes catégories de réseaux : le réseau à apprentissage supervisé, qui peut comparer son résultat avec une donnée en entrée, et le réseau à apprentissage non supervisé, qui ne connaît pas la réponse correcte, mais qui cherche à découvrir la structure sous-jacente des données par une recherche de corrélations entre les entrées afin de pouvoir organiser en catégories. Nous allons nous intéresser aux principes des réseaux à apprentissage supervisé avant de présenter les caractéristiques des réseaux non supervisé avec les cartes de Kohonen.

4 Présentation de l'outil de Data Mining

4.1 Principaux logiciels de Data Mining

4.1.1 SAS® Enterprise Miner

SAS® Enterprise Miner™ est une solution de Data Mining complète prposée par la société SAS. Elle repose sur une méthodologie éprouvée S.E.M.M.A. (Sample, Explore, Modify, Model, Assess). Elle offre une grande facilité de partage des résultats et d'exploitation des modèles.[3]

Cette plate-forme analytique permet de répondre aux besoins concrets des entreprises et les aide à optimiser leur stratégie avec une offre logicielle et un service de conseil et d'accompagnement reposant sur une expérience métier reconnue.

4.1.2 SPSS® Clementine

Clementine, l'atelier de Data Mining de SPSS, améliore l'efficacité de la relation client des entreprises en identifiant des groupes homogènes de consommateurs, des opportunités de ventes croisées, des cibles de nouveaux clients mais aussi en détectant des cas de fraudes. Clementine est fourni avec des applications prêtes pour analyser le comportement des internautes (Web Mining), calculer la valeur client, prévoir l'attrition, ou détecter des fraudes.[4]

4.1.3 Microsoft® SQL Server Analysis Services

Microsoft SQL Server 2005 Analysis Services fournit des fonctions OLAP et d'exploration de données pour les applications décisionnelles. Analysis Services prend en charge OLAP en permettant de concevoir, de créer et de gérer des structures multidimensionnelles qui contiennent des données agrégées provenant d'autres sources de données, telles que des bases de données relationnelles. Pour les applications d'exploration de données, Analysis Services permet de concevoir, de créer et de visualiser des modèles d'exploration de données créés à partir d'autres sources de données en utilisant un large éventail d'algorithmes d'exploration de données standard.[5]

4.2 Choix de l'outil de Data Mining

L'outil de Data Mining choisi pour ce projet est SQL Server Analysis Services car ce logiciel est disponible au SIMAV.

4.3 Les types de données

Ce chapitre est repris d'Internet [6].
Avant de traiter un modèle d'exploration de données dans SSAS, nous pouvons définir les types de données des colonnes de la structure d'exploration de données sur laquelle est basé le modèle. Analysis Services prend en charge les types de données suivants pour les colonnes de structure d'exploration de données :

[3] http://www.sas.com/offices/europe/france/software/documents/brochure_em.pdf
[4] http://www.spss.com/fr/clementine/
[5] http://technet.microsoft.com/fr-fr/library/ms175609.aspx
[6] http://technet.microsoft.com/fr-fr/library/ms174796.aspx

Type de données	Type de contenu
Text	Discrete, Discretized, Sequence
Long	Continuous, Cyclical, Discrete, Discretized, Key Sequence, Key Time, Ordered, Sequence, Time
Boolean	Discrete
Double	Continuous, Cyclique, Discrete, Discretized, Key Sequence, Key Time, Ordered, Sequence, Time
Date	Continuous, Discrete, Discretized, Key Time

Tableau 4-1 : Les types de données et leur type de contenu

Chacun de ces types de données prend en charge un ou plusieurs types de contenu que nous pouvons utiliser pour définir encore plus précisément les données des colonnes. Le Tableau 4-1 identifie les types de contenu pris en charge par chaque type de données.

4.4 Les types de contenu

Ce chapitre est repris d'Internet[7]*.*

Dans Microsoft SQL Server 2005 Analysis Services, nous pouvons définir les types de données des colonnes d'une structure d'exploration de données pour influer sur la manière dont les algorithmes traitent les données de ces colonnes lorsque nous créons des modèles d'exploration de données. Toutefois, la définition des types de données des colonnes donne uniquement aux algorithmes des informations sur le type de données figurant dans les colonnes et ne fournit aucune information sur le comportement de ces données. Pour cette raison, chaque type de données d'exploration de données dans Analysis Services prend en charge un ou plusieurs types de contenu que nous pouvons utiliser pour décrire le comportement du contenu des colonnes. Par exemple, si le contenu d'une colonne se répète selon un intervalle de temps spécifique, par exemple les jours de la semaine, nous pouvons définir le type de contenu de cette colonne comme étant cyclique.

La liste suivante décrit les types de contenu dans Analysis Services et identifie les types de données qui prennent en charge chacun de ces types. En plus des types de contenu répertoriés ici, nous pouvons utiliser des colonnes classifiées pour définir les types de contenu de certains types de données.

Type de contenu	Définition
DISCRETE	La colonne contient des valeurs discrètes. Par exemple, une colonne de genre (masculin/féminin) est une colonne d'attributs discrète typique, en ce sens que les données représentent un nombre fini et déterminé de catégories de genre. Les valeurs d'une colonne d'attributs discrète n'impliquent pas de tri des données, même si ce sont des valeurs numériques ; les valeurs sont clairement séparées, sans possibilité de valeurs fractionnaires. Les indicatifs téléphoniques sont un bon exemple de données discrètes numériques.
CONTINUOUS	La colonne contient des valeurs qui représentent un jeu continu de données numériques. À la différence d'une colonne discrète, qui représente des données finies et déterminées, une colonne continue représente des données de mesure et les données peuvent contenir un nombre infini de valeurs fractionnaires. Une colonne de revenus est un exemple de colonne d'attributs continue.

[7] http://technet.microsoft.com/fr-fr/library/ms174572.aspx

Type de contenu	Définition
DISCRETIZED	La colonne contient des valeurs qui représentent des groupes, ou compartiments, de valeurs dérivés d'une colonne continue. Les compartiments sont traités comme des valeurs discrètes et ordonnées.
KEY	La colonne identifie de manière unique une ligne.
KEY SEQUENCE	La colonne est un type de clé spécifique où les valeurs représentent une séquence d'événements. Les valeurs sont ordonnées et n'ont pas besoin d'être séparées par une distance égale. Ce type de contenu est pris en charge par les types de données suivants : Double, Long, Texte et Date.
KEY TIME	La colonne est un type de clé spécifique où les valeurs représentent des valeurs qui sont ordonnées et qui se produisent sur une période de temps donnée. Ce type de contenu est pris en charge par les types de données suivants : Double, Long et Date.
ORDERED	La colonne contient des valeurs qui définissent un jeu ordonné. Cependant, le jeu ordonné n'implique aucune relation de distance ou d'importance entre les valeurs du jeu. Par exemple, si une colonne d'attributs ordonnée contient des informations sur des niveaux de compétence classés de un à cinq, ceci n'implique pas une relation de distance entre les niveaux de compétence ; un niveau de compétence de valeur cinq n'est pas forcément cinq fois meilleur qu'un niveau de compétence de valeur un. Les colonnes d'attributs ordonnées sont considérées comme discrètes en termes de type de contenu.
CYCLICAL	La colonne contient des valeurs qui représentent un jeu ordonné cyclique. Par exemple, les jours numérotés de la semaine constituent un jeu ordonné cyclique car le jour numéro un suit le jour numéro sept. Les colonnes cycliques sont considérées comme ordonnées et discrètes en termes de type de contenu.

Tableau 4-2

4.5 Les algorithmes inclus dans SSAS

4.5.1 Microsoft Decision Trees (MDT)

Ce chapitre est repris d'Internet[8].

Description

L'algorithme MDT est un algorithme de classification et de régression fourni par Microsoft SQL Server 2005 Analysis Services et utilisé pour la modélisation prédictive d'attributs discrets et continus.

Pour les attributs discrets, l'algorithme effectue des prévisions en fonction des relations entre les colonnes d'entrée d'un dataset. Il utilise les valeurs ou les états de ces colonnes pour

[8] http://technet.microsoft.com/fr-fr/library/ms175312.aspx

prévoir les états d'une colonne désignée comme prévisible. En particulier, l'algorithme identifie les colonnes d'entrée en corrélation avec la colonne prévisible. Par exemple, dans un scénario conçu pour prévoir quels clients sont susceptibles d'acheter un vélo, si neuf jeunes clients sur dix achètent un vélo, alors que seulement deux clients plus âgés sur dix le font, l'algorithme déduit que l'âge est un bon facteur de prévision d'achat de vélo. L'arbre de décision effectue des prévisions en fonction de cette tendance vers une issue particulière.

Pour les attributs continus, l'algorithme utilise la régression linéaire pour déterminer où un arbre de décision se divise.

Si plusieurs colonnes sont définies comme prévisibles ou si les données d'entrée contiennent une table imbriquée définie comme prévisible, l'algorithme génère un arbre de décision distinct pour chaque colonne prévisible.

Fonctionnement
L'algorithme MDT crée un modèle d'exploration de données en créant une série de divisions, également appelées nœuds, dans l'arbre. L'algorithme ajoute un nœud au modèle chaque fois qu'une colonne d'entrée en corrélation significative avec la colonne prévisible est détectée. La manière dont l'algorithme détermine une division diffère selon qu'il prévoit une colonne continue ou une colonne discrète.

Prévision de colonnes discrètes
La manière dont l'algorithme MDT génère un arbre pour une colonne prévisible discrète peut être illustrée à l'aide d'un histogramme.

Prévision de colonnes continues
Lorsque l'algorithme MDT génère un arbre en fonction d'une colonne prévisible continue, chaque nœud contient une formule de régression. Une division apparaît à un point de non-linéarité dans la formule de régression.

Fonctions

Nom	Description
IsDescendant	Indique si le nœud actif descend du nœud spécifié.
IsInNode	Indique si le nœud spécifié contient le cas courant.
PredictAdjustedProbability	Retourne la probabilité ajustée d'une date spécifiée.
PredictAssociation	Prévoit une appartenance associative.
PredictHistogram	Retourne une table qui représente un histogramme de la prévision d'une colonne donnée.
PredictNodeId	Retourne le Node_ID du nœud dans lequel le cas est classé.
PredictProbability	Retourne la probabilité pour une date spécifiée.
PredictStdev	Retourne l'écart-type prévu pour la colonne spécifiée.
PredictSupport	Retourne la valeur de support pour une date spécifiée.
PredictVariance	Retourne la variance d'une colonne spécifiée.

Paramètres

Nom	Description
MAXIMUM_INPUT_ATTRIBUTES	Spécifie le nombre d'attributs d'entrée que l'algorithme peut traiter avant d'appeler la sélection des fonctionnalités. Attribuez à ce paramètre la valeur 0 pour désactiver la sélection des fonctionnalités. La valeur par défaut est 255.

Nom	Description
MAXIMUM_OUTPUT_ATTRIBUTES	Spécifie le nombre d'attributs de sortie que l'algorithme peut traiter avant d'appeler la sélection des fonctionnalités. Attribuez à ce paramètre la valeur 0 pour désactiver la sélection des fonctionnalités. La valeur par défaut est 255.
SCORE_METHOD	Spécifie la méthode utilisée pour calculer le résultat de la division. Options disponibles : Entropie (1), Bayésien avec a priori K2 (2) ou Équivalent bayésien de Dirichlet avec a priori (3). La valeur par défaut est 3.
SPLIT_METHOD	Spécifie la méthode utilisée pour diviser le nœud. Options disponibles : Binaire (1), Complet (2) ou Les deux (3). La valeur par défaut est 3.
MINIMUM_SUPPORT	Spécifie le nombre minimal de cas feuilles requis pour générer une division dans l'arbre de décision. La valeur par défaut est 10.
COMPLEXITY_PENALTY	Contrôle la croissance de l'arbre de décision. Une valeur faible entraîne l'augmentation du nombre de divisions, alors qu'une valeur importante entraîne la diminution du nombre de divisions. La valeur par défaut dépend du nombre d'attributs pour un modèle particulier, comme cela est décrit dans la liste suivante : De 1 à 9 attributs, la valeur par défaut est égale à 0,5. De 10 à 99 attributs, la valeur par défaut est égale à 0,9. À partir de 100 attributs, la valeur par défaut est égale à 0,99.
FORCED_REGRESSOR	Force l'algorithme à utiliser les colonnes indiquées comme régresseurs, quelle que soit leur importance, telle que calculée par l'algorithme. Ce paramètre est utilisé uniquement pour les arbres de décision qui prévoient un attribut continu.

4.5.2 Microsoft Naive Bayes (MNB)

Ce chapitre est repris d'Internet[9]*.*

L'algorithme MNB est un algorithme de classification fourni par Microsoft SQL Server 2005 Analysis Services qui est conçu pour la modélisation prédictive. Cet algorithme calcule la probabilité conditionnelle entre les colonnes d'entrée et les colonnes prévisibles, et suppose que les colonnes sont indépendantes. C'est en raison de cette supposition d'indépendance que l'algorithme s'appelle algorithme bayésien naïf (Naive Bayes). En effet, la supposition est souvent naïve étant donné que, en faisant cette supposition, l'algorithme ne prend pas en compte les dépendances qui peuvent exister.

Cet algorithme est informatiquement moins lourd que d'autres algorithmes Microsoft et est, par conséquent, utile pour générer rapidement des modèles d'exploration de données permettant de découvrir les relations entre les colonnes d'entrée et les colonnes prévisibles. Nous pouvons utiliser cet algorithme pour effectuer des explorations initiales de données et appliquer ensuite les résultats pour créer des modèles d'exploration de données supplémentaires avec d'autres algorithmes qui sont informatiquement plus lourds et plus précis.

Fonctionnement

L'algorithme MNB calcule la probabilité de tous les états de chaque colonne d'entrée, en fonction de chaque état possible de la colonne prévisible. Nous pouvons utiliser la Visionneuse de l'algorithme dans Business Intelligence Development Studio pour voir comment l'algorithme distribue les états.

Fonctions

Nom	Description
IsDescendant	Indique si le nœud actif descend du nœud spécifié.
PredictAdjustedProbability	Retourne la probabilité ajustée d'une date spécifiée.
PredictAssociation	Prévoit une appartenance associative.
PredictHistogram	Retourne une table qui représente un histogramme de la prévision d'une colonne donnée.
PredictNodeId	Retourne le Node_ID du nœud dans lequel le cas est classé.
PredictProbability	Retourne la probabilité pour une date spécifiée.
PredictSupport	Retourne la valeur de support pour une date spécifiée.

Paramètres

Nom	Description
MAXIMUM_INPUT_ATTRIBUTES	Spécifie le nombre maximal d'attributs d'entrée que l'algorithme peut traiter avant d'appeler la sélection des fonctionnalités. La valeur 0 désactive la sélection des fonctionnalités pour les attributs d'entrée. La valeur par défaut est 255.

[9] http://technet.microsoft.com/fr-fr/library/ms174806.aspx

Nom	Description
MAXIMUM_OUTPUT_ATTRIBUTES	Spécifie le nombre maximal d'attributs de sortie que l'algorithme peut traiter avant d'appeler la sélection des fonctionnalités. La valeur 0 désactive la sélection des fonctionnalités pour les attributs de sortie. La valeur par défaut est 255.
MINIMUM_DEPENDENCY_PROBABILITY	Spécifie la probabilité de dépendance minimale entre les attributs d'entrée et les attributs de sortie. Cette valeur sert à limiter la taille du contenu généré par l'algorithme. Cette propriété peut prendre une valeur comprise entre 0 et 1. Plus la valeur est grande, moins le nombre d'attributs dans le contenu du modèle est élevé. La valeur par défaut est 0,5.
MAXIMUM_STATES	Spécifie le nombre maximal d'états d'attribut que l'algorithme prend en charge. Si le nombre d'états d'un attribut est supérieur au nombre maximal d'états, l'algorithme sélectionne les états les plus fréquents pour cet attribut et traite les autres comme étant absents. La valeur par défaut est 100.

4.5.3 Clusters Microsoft

Ce chapitre est repris d'Internet[10].
L'algorithme Clusters Microsoft est un algorithme de segmentation fourni par Microsoft SQL Server 2005 Analysis Services. L'algorithme utilise des techniques itératives pour grouper les cas d'un jeu de données en clusters contenant des caractéristiques similaires. Ces groupements sont utiles pour l'exploration des données, l'identification d'anomalies dans les données et la création de prévisions.

Les modèles de clusters identifient des relations dans un jeu de données que nous ne pourrions peut-être pas déduire d'une observation informelle. Par exemple, nous pouvons déduire logiquement que les personnes qui se rendent à leur travail en vélo n'habitent généralement pas très loin de leur travail. Toutefois, l'algorithme peut trouver d'autres caractéristiques moins évidentes sur les personnes qui se rendent à leur travail en vélo. Dans le diagramme ci-dessous, le cluster A représente des données sur les personnes qui se rendent généralement en voiture à leur travail, tandis que le cluster B représente des données sur les personnes qui vont généralement en vélo à leur travail.

L'algorithme de clusters diffère des autres algorithmes d'exploration de données, tels que l'algorithme MDT, par le fait que nous n'avons pas à désigner de colonne prévisible pour être en mesure de générer un modèle de clusters. L'algorithme de clusters effectue l'apprentissage du modèle strictement à partir des relations qui existent dans les données et à partir des clusters que l'algorithme identifie.

Fonctionnement
L'algorithme Clusters Microsoft commence par identifier les relations qui existent dans un jeu de données et génère une série de clusters en fonction de ces relations. Un nuage de points représente une méthode utile pour représenter graphiquement la manière dont l'algorithme groupe les données.

Une fois les clusters définis, l'algorithme calcule comment les clusters représentent les groupements des points, puis tente de redéfinir les groupements pour créer des clusters qui représentent mieux les données. L'algorithme effectue des itérations sur ce processus jusqu'à ce qu'il ne puisse plus améliorer les résultats en redéfinissant les clusters.

L'algorithme Clusters Microsoft offre deux méthodes pour calculer comment les points sont adaptés aux clusters : EM (Expectation Maximization) et K-Means. Pour les clusters EM, l'algorithme utilise une méthode probabiliste pour déterminer la probabilité qu'un point de données existe dans un cluster. Pour K-Means, l'algorithme utilise une mesure de distance pour attribuer un point de données au cluster le plus proche.

Les colonnes dont l'utilisation est définie pour prévoir uniquement ne sont pas utilisées pour générer des clusters. Leurs distributions dans les clusters sont calculées après la création des clusters.

Fonctions

Nom	Description
Cluster	Retourne le cluster le plus susceptible de contenir le cas d'entrée.
ClusterProbability	Retourne la probabilité que le cas d'entrée appartient au cluster spécifié.

[10] http://technet.microsoft.com/fr-fr/library/ms174879.aspx

Nom	Description
IsDescendant	Indique si le nœud actif descend du nœud spécifié.
IsInNode	Indique si le nœud spécifié contient le cas courant.
PredictAdjustedProbability	Retourne la probabilité ajustée d'une date spécifiée.
PredictAssociation	Prévoit une appartenance associative.
PredictHistogram	Retourne une table qui représente un histogramme de la prévision d'une colonne donnée.
PredictCaseLikelihood	Uniquement utilisée avec les modèles de clustering. Cette fonction retourne la probabilité qu'un cas d'entrée rentre dans le modèle existant.
PredictNodeId	Retourne le Node_ID du nœud dans lequel le cas est classé.
PredictProbability	Retourne la probabilité pour une date spécifiée.
PredictStdev	Retourne l'écart-type prévu pour la colonne spécifiée.
PredictSupport	Retourne la valeur de support pour une date spécifiée.
PredictVariance	Retourne la variance d'une colonne spécifiée.

Paramètres

Nom	Description
CLUSTERING_METHOD	Spécifie la méthode de clustering que doit utiliser l'algorithme. Les méthodes de clustering disponibles sont les suivantes : EM dimensionnable (1), EM non dimensionnable (2), K-Means dimensionnable (3) et K-Means non dimensionnable(4). La valeur par défaut est 1.
CLUSTER_COUNT	Spécifie le nombre approximatif de clusters que l'algorithme doit générer. S'il est impossible de générer ce nombre approximatif de clusters à partir des données, l'algorithme génère autant de clusters que possible. Si CLUSTER_COUNT a la valeur 0, l'algorithme utilise des valeurs heuristiques pour déterminer de manière optimale le nombre de clusters à générer. La valeur par défaut est 10.
CLUSTER_SEED	Spécifie la valeur de départ utilisée pour générer de façon aléatoire des clusters pour le stade initial de construction d'un modèle. La valeur par défaut est 0.
MINIMUM_SUPPORT	Spécifie le nombre minimal de cas dans chaque cluster. La valeur par défaut est 1.
MODELLING_CARDINALITY	Spécifie le nombre d'exemples de modèles générés pendant le processus de clustering. La valeur par défaut est 10.

Nom	Description
STOPPING_TOLERANCE	Spécifie la valeur utilisée pour déterminer quand une convergence est atteinte et quand l'algorithme a terminé la construction du modèle. La convergence est atteinte lorsque le changement global dans les probabilités de clusters est inférieur au rapport du paramètre STOPPING_TOLERANCE divisé par la taille du modèle. La valeur par défaut est 10.
SAMPLE_SIZE	Spécifie le nombre de cas que l'algorithme utilise à chaque passage si l'une des méthodes de clustering dimensionnable est définie pour le paramètre CLUSTERING_METHOD. Si la valeur 0 est attribuée au paramètre SAMPLE_SIZE, le jeu de données complet est organisé en clusters en un seul passage. Cela risque de créer des problèmes de mémoire et de performances. La valeur par défaut est 50000.
MAXIMUM_INPUT_ATTRIBUTES	Spécifie le nombre maximal d'attributs d'entrée que l'algorithme peut traiter avant d'appeler la sélection des fonctionnalités. La valeur 0 spécifie qu'il n'y a pas de nombre maximal d'attributs. La valeur par défaut est 255.
MAXIMUM_STATES	Spécifie le nombre maximal d'états d'attribut que l'algorithme prend en charge. Si le nombre d'états d'un attribut est supérieur au nombre maximal d'états, l'algorithme emploie les états les plus utilisés de l'attribut et ignore les autres états. La valeur par défaut est 100.

4.5.4 Microsoft Neural Network (MNN)

Ce chapitre est repris d'Internet[11].
Dans Microsoft SQL Server 2005 Analysis Services, l'algorithme MNN crée des modèles d'exploration de données de classification et de régression en construisant un réseau perceptron multicouche de neurones. Similaire à l'algorithme MDT, l'algorithme MNN calcule les probabilités de chaque état possible de l'attribut d'entrée lorsque chaque état de l'attribut prévisible lui est fourni. Nous pouvons par la suite utiliser ces probabilités pour prédire le résultat de l'attribut prédit en fonction des attributs d'entrée.

Fonctionnement
L'algorithme MNN utilise un réseau perceptron multicouche, également appelé réseau à règle delta à rétro propagation, qui peut comporter jusqu'à trois couches de neurones, ou perceptrons : une couche d'entrée, une couche cachée facultative et une couche de sortie. Dans un réseau perceptron multicouche, chaque neurone reçoit une ou plusieurs entrées et produit une ou plusieurs sorties identiques. Chaque sortie est une fonction non linéaire simple de la somme des entrées dans le neurone. Les entrées sont uniquement propagées vers l'avant entre des nœuds de la couche d'entrée et des nœuds de la couche cachée avant de passer finalement à la couche de sortie ; il n'y a aucune connexion entre des neurones au sein d'une même couche. (Les entrées sont propagées vers l'avant entre des nœuds de la couche d'entrée et des nœuds de la couche de sortie s'il n'y a pas de couche cachée.) Les réseaux de neurones de type perceptron multicouche ne sont pas décrits en détail dans la présente documentation.

Un modèle d'exploration de données généré avec l'algorithme MNN peut contenir plusieurs réseaux, en fonction du nombre de colonnes utilisées soit pour l'entrée et la prédiction, soit uniquement pour la prédiction. Le nombre de réseaux d'un modèle d'exploration de données dépend du nombre d'états figurant dans les colonnes d'entrée et dans les colonnes prévisibles utilisées par ce modèle d'exploration de données.

Il existe trois types de neurones dans un réseau de neurones créé avec l'algorithme MNN :

- Neurones d'entrée
  Les neurones d'entrée fournissent les valeurs des attributs d'entrée du modèle d'exploration de données. Pour les attributs d'entrée discrets, un neurone d'entrée représente généralement un état unique de l'attribut d'entrée, y compris un état manquant. Par exemple, un attribut d'entrée binaire produit un nœud d'entrée qui décrit un état existant ou manquant, indiquant si une valeur existe ou non pour cet attribut. Une colonne booléenne utilisée en tant qu'attribut d'entrée génère trois neurones d'entrée : un neurone pour une valeur vraie, un neurone pour une valeur fausse et un neurone pour un état manquant ou existant. Un attribut d'entrée discret qui a plus de deux états génère un neurone d'entrée pour chaque état et un neurone d'entrée pour un état manquant ou existant. Un attribut d'entrée continu génère deux neurones d'entrée : un neurone pour un état manquant ou existant et un neurone pour la valeur de l'attribut continu lui-même. Les neurones d'entrée fournissent des entrées à un ou plusieurs neurones cachés.
- Neurones cachés
  Les neurones cachés reçoivent des entrées des neurones d'entrée et fournissent des sorties aux neurones de sortie.
- Neurones de sortie
  Les neurones de sortie représentent les valeurs des attributs prévisibles du modèle d'exploration de données. Pour les attributs d'entrée discrets, un neurone de sortie

[11] http://technet.microsoft.com/fr-fr/library/ms174941.aspx

représente généralement un état prédit unique d'un attribut prévisible, y compris un état manquant. Par exemple, un attribut prévisible binaire produit un nœud de sortie qui décrit un état existant ou manquant pour indiquer si une valeur existe ou non pour cet attribut. Une colonne booléenne utilisée en tant qu'attribut prévisible génère trois neurones de sortie : un neurone pour une valeur vraie, un neurone pour une valeur fausse et un neurone pour un état manquant ou existant. Un attribut prévisible discret qui a plus de deux états génère un neurone de sortie pour chaque état et un neurone de sortie pour un état manquant ou existant. Les colonnes prévisibles continues génèrent deux neurones de sortie : un neurone pour un état manquant ou existant et un neurone pour la valeur de la colonne continue elle-même. Si plus de 500 neurones de sortie sont générés par l'examen de l'ensemble des colonnes prévisibles, Analysis Services génère un nouveau réseau dans le modèle d'exploration de données pour représenter les neurones de sortie supplémentaires.

Un neurone reçoit plusieurs entrées : avec les neurones d'entrée, un neurone reçoit des entrées à partir des données d'origine ; avec les neurones cachés et les neurones de sortie, un neurone reçoit des entrées provenant de la sortie d'autres neurones du réseau de neurones. Les entrées établissent des relations entre les neurones et ces relations servent de chemin d'analyse pour un ensemble de cas spécifique.

Chaque entrée est dotée d'une valeur, appelée le poids, qui décrit la pertinence ou l'importance d'une entrée donnée par rapport au neurone caché ou au neurone de sortie. Plus le poids attribué à une entrée est grand, plus la valeur de cette entrée est pertinente ou importante pour le neurone qui reçoit lorsque l'algorithme détermine si cette entrée classifie un cas spécifique. Notons également que le poids peut être négatif, ce qui implique que l'entrée peut désactiver un neurone spécifique au lieu de l'activer. La valeur de l'entrée est multipliée par le poids pour accentuer son importance pour un neurone spécifique. (Si le poids est négatif, la valeur de l'entrée est multipliée par le poids pour désaccentuer son importance.)

En conséquence, chaque neurone est doté d'une fonction non linéaire simple, appelée fonction d'activation, qui décrit la pertinence ou l'importance d'un neurone donné par rapport à la couche d'un réseau de neurones. Les neurones cachés utilisent une fonction tangente hyperbolique pour leur fonction d'activation, tandis que les neurones de sortie utilisent une fonction sigmoïde pour leur fonction d'activation. Ces deux fonctions sont des fonctions continues non linéaires qui permettent au réseau de neurones de modéliser les relations non linéaires entre les neurones d'entrée et les neurones de sortie.

Fonctions

Nom	Description
IsDescendant	Indique si le nœud actif descend du nœud spécifié.
PredictAdjustedProbability	Retourne la probabilité ajustée d'une date spécifiée.
PredictHistogram	Retourne une table qui représente un histogramme de la prévision d'une colonne donnée.
PredictProbability	Retourne la probabilité pour une date spécifiée.
PredictStdev	Retourne l'écart-type prévu pour la colonne spécifiée.
PredictSupport	Retourne la valeur de support pour une date spécifiée.
PredictVariance	Retourne la variance d'une colonne spécifiée.

Paramètres

Nom	Description
HIDDEN_NODE_RATIO	Spécifie le taux de neurones cachés par rapport aux neurones d'entrée et de sortie. La formule suivante détermine le nombre initial de neurones de la couche cachée : HIDDEN_NODE_RATIO * SQRT(Total input neurons * Total output neurons) La valeur par défaut est 4,0.
HOLDOUT_PERCENTAGE	Spécifie le pourcentage de cas extraits des données d'apprentissage pour calculer l'erreur d'exclusion, qui constitue l'un des critères d'arrêt pendant l'apprentissage du modèle d'exploration de données. La valeur par défaut est 30.
HOLDOUT_SEED	Spécifie un nombre qui est utilisé en tant que valeur de départ du générateur de nombres pseudo-aléatoires lorsque l'algorithme détermine de façon aléatoire les données d'exclusion. Si ce paramètre a la valeur 0, l'algorithme génère la valeur de départ en fonction du nom du modèle d'exploration de données, afin de garantir que le contenu du modèle reste inchangé pendant le retraitement. La valeur par défaut est 0.
MAXIMUM_INPUT_ATTRIBUTES	Détermine le nombre maximal d'attributs d'entrée qui peuvent être fournis à l'algorithme et au-delà duquel la sélection des fonctionnalités est utilisée. La valeur 0 désactive la sélection des fonctionnalités pour les attributs d'entrée. La valeur par défaut est 255.
MAXIMUM_OUTPUT_ATTRIBUTES	Détermine le nombre maximal d'attributs de sortie qui peuvent être fournis à l'algorithme et au-delà duquel la sélection des fonctionnalités est utilisée. La valeur 0 désactive la sélection des fonctionnalités pour les attributs de sortie. La valeur par défaut est 255.

Nom	Description
MAXIMUM_STATES	Spécifie le nombre maximal d'états discrets par attribut qui est pris en charge par l'algorithme. Si le nombre d'états d'un attribut spécifique est supérieur au nombre spécifié pour ce paramètre, l'algorithme sélectionne les états les plus fréquents pour cet attribut et traite les autres comme étant absents. La valeur par défaut est 100.
SAMPLE_SIZE	Spécifie le nombre de cas à utiliser pour l'apprentissage du modèle. L'algorithme utilise soit ce nombre, soit le pourcentage du nombre total de cas qui n'est pas inclus dans les données d'exclusion conformément au paramètre HOLDOUT_PERCENTAGE : c'est la plus petite valeur qui est retenue. En d'autres termes, si HOLDOUT_PERCENTAGE a la valeur 30, l'algorithme utilisera soit la valeur de ce paramètre, soit une valeur égale à 70 % du nombre total de cas, en prenant la plus petite valeur des deux. La valeur par défaut est 10000.

4.5.5 Microsoft Time Series (MTS)

Ce chapitre est repris d'Internet[12]*.*

L'algorithme MTS est un algorithme de régression fourni par Microsoft SQL Server 2005 Analysis Services qui est conçu pour la création de modèles d'exploration de données permettant de prédire des colonnes continues, telles que des ventes de produits, dans un scénario de prévision. Alors que d'autres algorithmes Microsoft créent des modèles, tels que les modèles d'arbre de décision, qui se basent sur des colonnes d'entrée pour prédire la colonne prévisible, la prédiction dans un modèle de série chronologique est uniquement basée sur les tendances que l'algorithme dégage dans le dataset d'origine pendant la création du modèle.

Fonctionnement

L'algorithme MDT entraîne un modèle à l'aide d'un arbre de décision autorégressif. Chaque modèle contient une colonne de temps clé qui définit les tranches de temps que le modèle va définir. L'algorithme associe un nombre variable d'éléments passés à chaque élément actuel prédit.

Fonctions

Nom	Description
Lag	Retourne la tranche de temps entre la date du cas en cours et la dernière date de l'ensemble d'apprentissage.
PredictNodeId	Retourne le Node_ID du nœud dans lequel le cas est classé.
PredictStdev	Retourne le Node_ID du nœud dans lequel le cas est classé.
PredictTimeSeries	Retourne des valeurs de prévision dans le future ou historiques pour les données de séries chronologiques. Les données de séries chronologiques étant continues, elles ne peuvent être stockées ni dans une table imbriquée ni dans une table de cas. La fonction PredictTimeSeries retourne toujours une table imbriquée.
PredictVariance	Retourne la variance d'une colonne spécifiée.

Paramètres

Nom	Description
MINIMUM_SUPPORT	Spécifie le nombre minimal de tranches de temps qui sont requises pour générer un fractionnement dans chaque arbre de série chronologique. La valeur par défaut est 10.
COMPLEXITY_PENALTY	Contrôle la croissance de l'arbre de décision. La diminution de cette valeur augmente la probabilité d'une division. L'augmentation de cette valeur diminue la probabilité d'une division. La valeur par défaut est 0,1.

[12] http://technet.microsoft.com/fr-fr/library/ms174923.aspx

Nom	Description
PERIODICITY_HINT	Fournit à l'algorithme une indication de la périodicité des données. Par exemple, si les ventes varient chaque année et que l'unité de mesure de la série est le mois, la périodicité est égale à 12. Ce paramètre s'affiche sous la forme {n [, n]}, où n est un nombre positif. Le n entre crochets [] est facultatif et peut être répété aussi souvent que nécessaire. La valeur par défaut est {1}.
MISSING_VALUE_SUBSTITUTION	Spécifie la méthode employée pour combler les vides dans les données d'historique. Par défaut, les vides et les extrémités irréguliers ne sont pas autorisés dans les données. Les méthodes disponibles pour combler les vides et les extrémités irréguliers sont les suivantes : par Valeur précédente, par Valeur moyenne ou par constante numérique spécifique.
AUTO_DETECT_PERIODICITY	Spécifie une valeur numérique comprise entre 0 et 1 utilisée pour détecter la périodicité. Une valeur proche de 1 favorise la découverte de nombreux modèles quasi-périodiques et la génération automatique d'indications de périodicité. Le traitement d'un grand nombre d'indications de périodicité est susceptible d'allonger de façon significative les durées d'apprentissage des modèles et de produire des modèles plus précis. Si la valeur est proche de 0, la périodicité n'est détectée que pour les données fortement périodiques. La valeur par défaut est 0,6.
HISTORIC_MODEL_COUNT	Spécifie le nombre de modèles historiques qui seront générés. La valeur par défaut est 1.
HISTORICAL_MODEL_GAP	Spécifie le décalage dans le temps entre deux modèles historiques successifs. Par exemple, la valeur g produit des modèles historiques générés pour des données tronquées par tranches de temps à des intervalles de g, 2*g, 3*g et ainsi de suite. La valeur par défaut est 10.

4.5.6 Microsoft Sequence Clustering (MSC)

Ce chapitre est repris d'Internet[13].
L'algorithme MSC est un algorithme d'analyse de séquence fourni par Microsoft SQL Server 2005 Analysis Services. Cet algorithme vous permet d'explorer des données qui contiennent des événements qui peuvent être liés en suivant des chemins ou des séquences. L'algorithme recherche les séquences les plus communes en groupant, ou en regroupant en clusters, les séquences identiques.

Fonctionnement
L'algorithme utilise la méthode de clustering EM pour identifier les clusters et leurs séquences. En particulier, l'algorithme utilise une méthode probabiliste pour déterminer la probabilité qu'un point de données existe dans un cluster. Pour obtenir une description de l'utilisation de cette méthode de clustering dans l'algorithme Clusters Microsoft, consulter Algorithme Clusters Microsoft.

Une des colonnes d'entrée que l'algorithme MSC utilise est une table imbriquée qui contient des séquences de données. Ces données représentent une série de transitions d'état de cas individuels dans un jeu de données, telles que des achats de produits ou des clics Web. Pour déterminer les colonnes de séquence à traiter comme colonnes d'entrée pour le clustering, l'algorithme mesure les différences, ou les distances, entre toutes les séquences possibles dans le jeu de données. Une fois que l'algorithme a mesuré ces distances, il peut utiliser la colonne de séquence comme entrée pour la méthode EM de clustering.

Fonctions

Nom	Description
Cluster	Retourne le cluster le plus susceptible de contenir le cas d'entrée.
ClusterProbability	Retourne la probabilité que le cas d'entrée appartient au cluster spécifié.
IsDescendant	Indique si le nœud actif descend du nœud spécifié.
IsInNode	Indique si le nœud spécifié contient le cas courant.
PredictAdjustedProbability	Retourne la probabilité ajustée d'une date spécifiée.
PredictAssociation	Prévoit une appartenance associative.
PredictCaseLikelihood	Uniquement utilisée avec les modèles de clustering. Cette fonction retourne la probabilité qu'un cas d'entrée rentre dans le modèle existant.
PredictHistogram	Retourne une table qui représente un histogramme de la prévision d'une colonne donnée.
PredictNodeId	Retourne le Node_ID du nœud dans lequel le cas est classé.
PredictProbability	Retourne la probabilité pour une date spécifiée.
PredictSequence	Prédit les valeurs de séquence pour un ensemble spécifié de données de séquence.
PredictStdev	Retourne l'écart-type prévu pour la colonne spécifiée.
PredictSupport	Retourne la valeur de support pour une date spécifiée.
PredictVariance	Retourne la variance d'une colonne spécifiée.

[13] http://technet.microsoft.com/fr-fr/library/ms175462.aspx

Paramètres

Nom	Description
CLUSTER_COUNT	Spécifie le nombre approximatif de clusters que l'algorithme doit générer. S'il est impossible de générer ce nombre approximatif de clusters à partir des données, l'algorithme génère autant de clusters que possible. Si CLUSTER_COUNT a la valeur 0, l'algorithme utilise des valeurs heuristiques pour déterminer de manière optimale le nombre de clusters à générer. La valeur par défaut est 10.
MINIMUM_SUPPORT	Spécifie le nombre minimal de cas dans chaque cluster. La valeur par défaut est 10.
MAXIMUM_SEQUENCE_STATES	Spécifie le nombre maximal d'états qu'une séquence peut avoir. Si cette valeur est supérieure à 100, l'algorithme peut créer un modèle qui ne fournit pas d'informations significatives. La valeur par défaut est 64.
MAXIMUM_STATES	Spécifie le nombre maximal d'états d'attribut que l'algorithme prend en charge. Si le nombre d'états d'un attribut est supérieur au nombre maximal d'états, l'algorithme emploie les états les plus utilisés de l'attribut et ignore les autres états. La valeur par défaut est 100.

4.5.7 Algorithme Microsoft Association

Ce chapitre est repris d'Internet[14].

L'algorithme Microsoft Association est un algorithme d'association fourni par Microsoft SQL Server 2005 Analysis Services, qui est utile pour les moteurs de recommandation. Un moteur de recommandation recommande des produits aux clients en se basant sur les éléments qu'ils ont déjà achetés ou pour lesquels ils ont manifesté un intérêt. L'algorithme Microsoft Association est utile également pour l'analyse d'un panier d'achats.

Fonctionnement

L'algorithme Microsoft Association parcourt un jeu de données pour trouver les éléments qui apparaissent ensemble dans un cas. L'algorithme groupe alors en jeux d'éléments tous les éléments associés qui apparaissent un nombre de fois au moins égal au nombre de cas spécifiés par le paramètre MINIMUM_SUPPORT. Par exemple, un jeu d'éléments peut être « Mountain 200=Existing, Sport 100=Existing » et peut avoir une prise en charge de 710. L'algorithme génère alors des règles à partir des jeux d'éléments. Ces règles sont utilisées pour prévoir la présence d'un élément dans la base de données, en fonction de la présence d'autres éléments spécifiques que l'algorithme identifie comme importants. Par exemple, une règle peut être « if Touring 1000=existing and Road bottle cage=existing, then Water bottle=existing » et peut avoir une probabilité de 0,812. Dans cet exemple, l'algorithme identifie le fait que la présence dans le panier d'un pneu Touring 1000 et d'un porte-bidon indique qu'un bidon d'eau pourrait également probablement se trouver dans le panier.

Fonctions

Nom	Description
IsDescendant	Indique si le nœud actif descend du nœud spécifié.
IsInNode	Indique si le nœud spécifié contient le cas courant.
PredictAdjustedProbability	Retourne la probabilité ajustée d'une date spécifiée.
PredictAssociation	Prévoit une appartenance associative.
PredictHistogram	Retourne une table qui représente un histogramme de la prévision d'une colonne donnée.
PredictNodeId	Retourne le Node_ID du nœud dans lequel le cas est classé.
PredictProbability	Retourne la probabilité pour une date spécifiée.
PredictSupport	Retourne la valeur de support pour une date spécifiée.

[14] http://technet.microsoft.com/fr-fr/library/ms174916.aspx

Paramètres

Nom	Description
MINIMUM_SUPPORT	Spécifie le nombre minimal de cas qui doivent contenir le jeu d'éléments avant que l'algorithme génère une règle. Attribuer à ce paramètre une valeur inférieure à 1 spécifie ce nombre minimal de cas comme un pourcentage du nombre total de cas. Attribuer à ce paramètre une valeur entière supérieure à 1 spécifie le nombre minimal de cas comme le nombre absolu de cas qui doivent contenir le jeu d'éléments. L'algorithme peut augmenter la valeur de ce paramètre si la mémoire est limitée. La valeur par défaut est 0,03.
MAXIMUM_SUPPORT	Spécifie le nombre maximal de cas pour lesquels un jeu d'éléments peut bénéficier d'un support. Une valeur inférieure à 1 représente un pourcentage du nombre total de cas. Une valeur supérieure à 1 représente le nombre absolu de cas qui peuvent contenir le jeu d'éléments. La valeur par défaut est 1.
MINIMUM_ITEMSET_SIZE	Spécifie le nombre minimal d'éléments autorisés dans un jeu d'éléments. La valeur par défaut est 1.
MAXIMUM_ITEMSET_SIZE	Spécifie le nombre maximal d'éléments autorisés dans un jeu d'éléments. Une valeur de 0 spécifie qu'il n'y a pas de limite quant à la taille du jeu d'éléments. La valeur par défaut est 3.
MAXIMUM_ITEMSET_COUNT	Spécifie le nombre maximal de jeux d'éléments à produire. Si aucun nombre n'est spécifié, l'algorithme génère tous les jeux d'éléments possibles. La valeur par défaut est 200000.
MINIMUM_PROBABILITY	Spécifie la probabilité minimale qu'une règle ait la valeur True. Par exemple, la valeur 0,5 spécifie qu'aucune règle présentant une probabilité inférieure à 50 % n'est générée. La valeur par défaut est 0,4.
OPTIMIZED_PREDICTION_COUNT	Définit le nombre d'éléments à mettre en cache ou à optimiser pour une prédiction.

5 Description détaillée des fonctionnalités de PrediRec

5.1.1 Use Case

PrediRec – Utilisation Multi-Case

Dans ce scénario (Figure 5-1), l'utilisateur peut choisir le modèle d'exploration de données, les cas sortis et non codés ainsi que les cas non sortis.
Une fois les cas choisis, l'utilisateur peut les estimer et, si besoin, les exporter dans MS Excel.

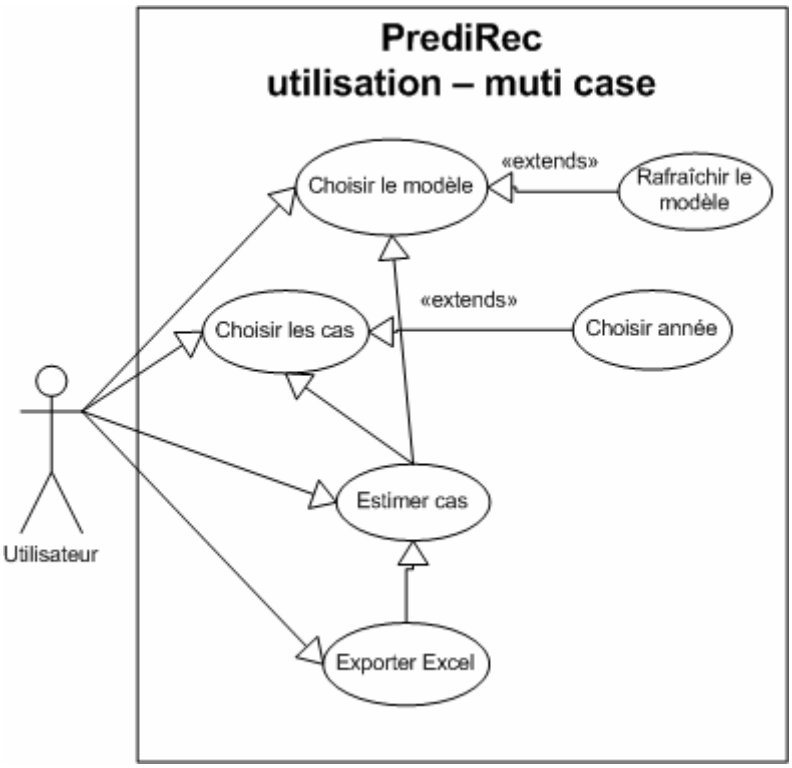


Figure 5-1 : Use case : Multi-Case

PrediRec – Utilisation Single-Case
Dans ce scénario (Figure 5-2), l'utilisateur peut créer un cas fictif en renseignant chaque variable nécessaire à sa simulation et le soumettre au modèle d'analyse de données choisi.

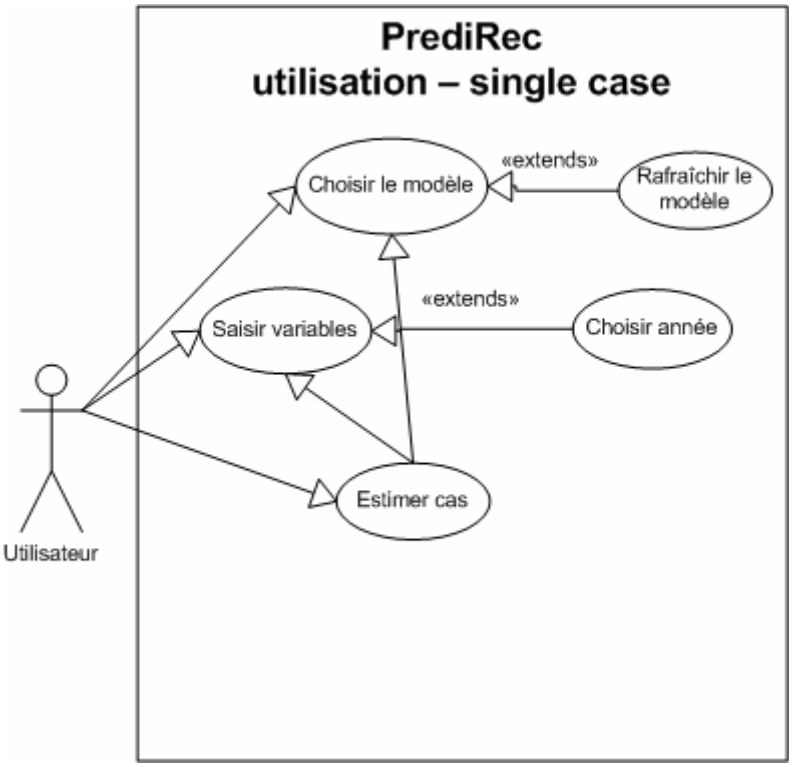


Figure 5-2 : Use case : Single case

5.1.2 Description du Use Case : PrediRec – Utilisation Multi Case

Etant donné que la principale fonctionnalité du côté utilisateur est la prédiction des recettes liées aux cas non codés, voici une description plus précise de ce scénario.

L'utilisateur, une fois connecté sur le système PrediRec, a la possibilité de choisir la variable à prédire : soit le « CW pondéré », soit le « Total Facture ».
Ensuite, il doit sélectionner l'année de référence du modèle. Nous laissons à l'utilisateur cette possibilité, car un modèle de « CW pondéré » peut être utilisé d'une année à l'autre et cela tant que la version des APDRG reste la même. Par contre, un modèle de « Total Facture » ne doit pas être utilisé d'une année à l'autre, sauf à titre indicatif.

L'application permet de choisir parmi tous les cas non codés selon les critères suivants :
- l'année de sortie dans une liste déroulante
- les cas non sortis de l'hôpital via une case à cocher

Nb. L'utilisateur peut cumuler les deux options ci-dessus.

Une fois ces choix effectués, PrediRec effectue une requête sur la base de données source et en extrait une liste de cas qu'il présente sous forme de tableau.
A partir de cette liste, l'utilisateur a la possibilité de choisir lesquels des cas présentés il désire estimer. Par défaut, tous les cas sont sélectionnés.
Une fois les cas choisis, l'utilisateur peut lancer la simulation.
A la fin de celle-ci, le système affiche pour chaque cas la valeur estimée et offre la possibilité de les exporter vers MS Excel.

5.1.3 Diagrammes de séquences

Ci-dessous (Figure 5-3), le diagramme de séquence de l'Use-Case PrediRec - MultiValue

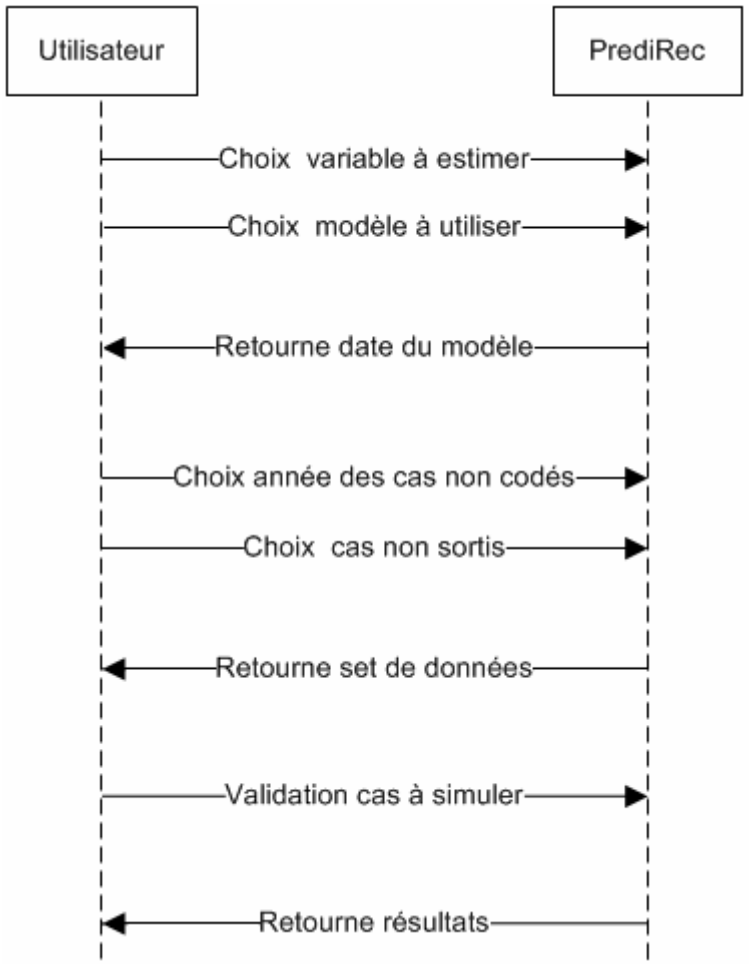


Figure 5-3 :Diagramme de séquence

5.1.4 Interfaces utilisateurs

5.1.4.1 Interface Multi-Case

Voici l'interface du site Internet pour la simulation de plusieurs cas (Figure 5-4) :

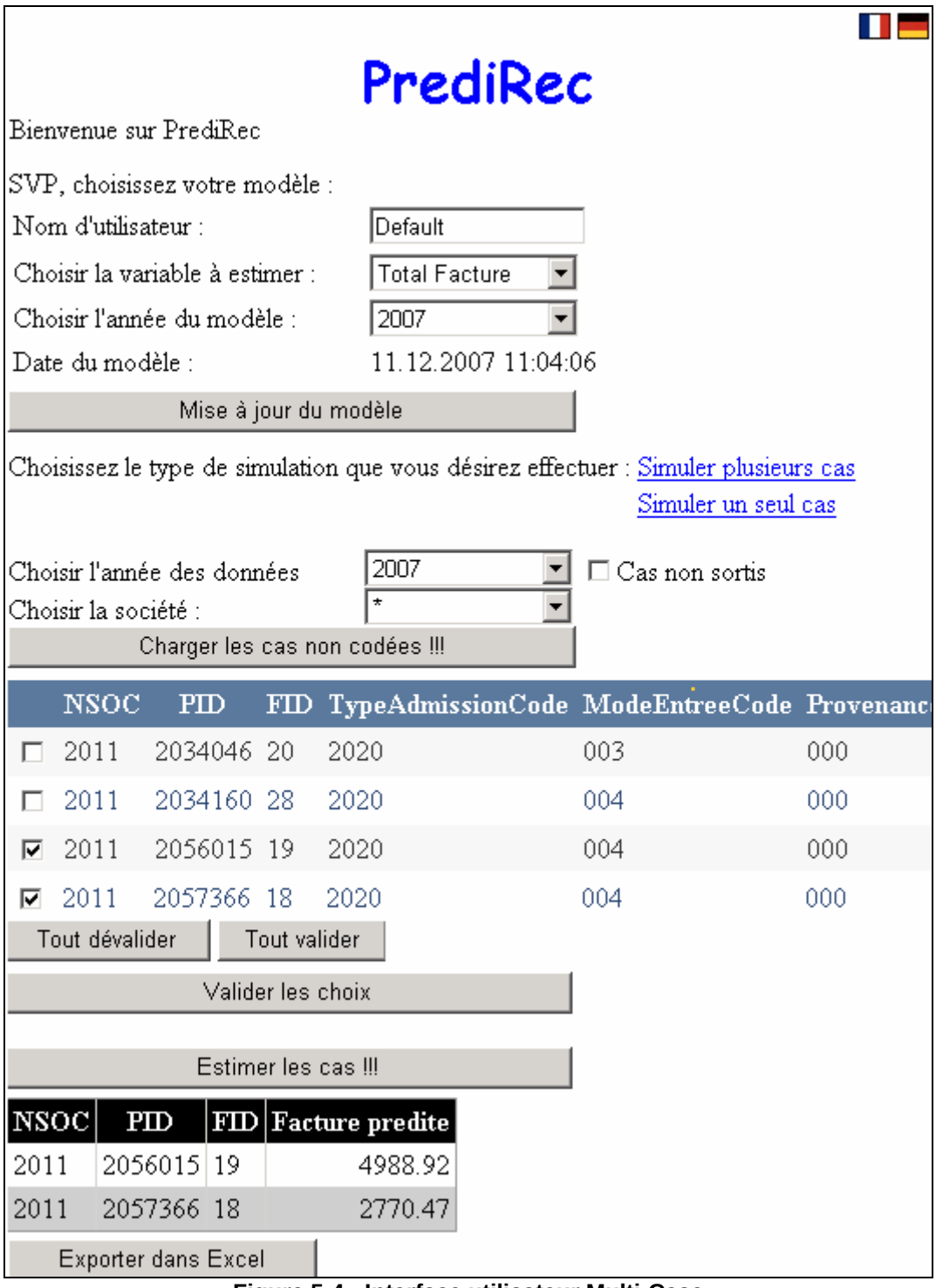


Figure 5-4 : Interface utilisateur Multi-Case

L'utilisation de cette page est très intuitive et nécessite, mise à part pour la saisie du nom d'utilisateur, uniquement l'utilisation de la souris.

5.1.4.2 Interface Single-Case

Voici l'interface pour la spécification d'un cas unique :

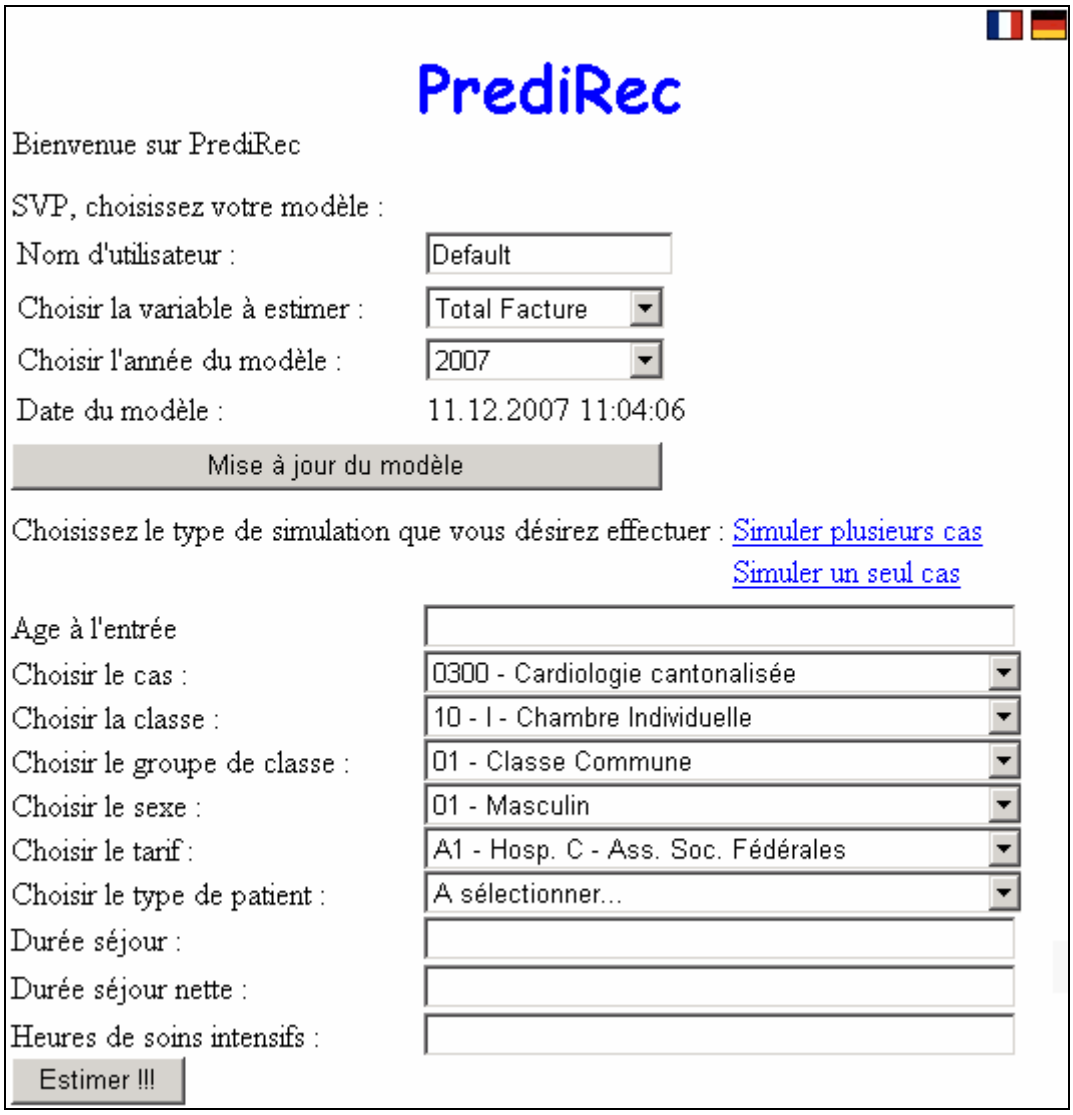


Figure 5-5 : Interface utilisateur Single-Case

L'utilisation de cette page est un peu moins intuitive que la page « Multi-Case » (5.1.4.1 - Interface Multi-Case) car elle demande, de la part de l'utilisateur, des informations précises.

6 Définition des modèles d'analyse de PrediRec

Etant donné que nous avons défini plus haut la méthodologie d'un projet de Data Mining (« 3.2 - Méthodologies du Data Mining »), passons de la théorie à la pratique.

6.1 Définition du problème

Les comptables ont sollicité le SIMAV afin qu'il leur mette à disposition un outil permettant d'approximer au mieux les recettes de cas non codés (un cas correspond à un séjour d'un patient hospitalisé).

Une séance avec M. Werlen, Chef de la division « Finances et Controlling » du SZO, fut nécessaire afin de correctement définir le cadre du projet et aussi pour définir ce qu'est un cas non codé à estimer :

- Est-ce uniquement les cas sortis de l'hôpital à un moment X et qui ne sont pas encore codés et facturés ?
- Est-ce que les cas toujours présents à l'hôpital doivent aussi être estimés ?

Il en résulta que nous devons effectuer une estimation des cas sortis non codés et aussi des cas présents dans les hôpitaux, tout en sachant que les estimations effectuées sur ces derniers ne peuvent être utilisées qu'à titre informatif.
En effet, il est évident que les coûts liés à une hospitalisation sont dépendants de la durée de séjour et que celle-ci peut être calculée uniquement à la sortie du patient de l'hôpital.

6.2 Choix des variables à estimer

Nous proposons à l'utilisateur deux méthodes différentes pour estimer les cas non codés.

La première méthode permet d'estimer directement le « total facturé » au patient. L'avantage de cette méthode est qu'elle permet d'approximer plus correctement les recettes liées aux cas non codés. L'inconvénient est, par contre, que le modèle d'analyse a besoin de beaucoup de cas d'apprentissage et qu'il n'est pas utilisable, sauf à titre indicatif, d'une année à l'autre.

La deuxième méthode permet d'estimer le « Cost-Weight pondéré » lié aux cas non codés. L'avantage de celle-ci est que tant que la version des APDRGs ne change pas d'une année à l'autre, le modèle peut être utilisé sans devoir être au préalable adapté. L'inconvénient par contre, est le fait que l'estimation en résultant est moins précise que celle effectuée avec la méthode précédente.

Il est nécessaire aussi de savoir que le «Total Facture » est égal au « CW pondéré » multiplié par la valeur du point, valeur qui est réévaluée chaque année.

L'application PrediRec permet donc à l'utilisateur de choisir l'une ou l'autre variable pour effectuer l'estimation de ces cas non codés.

6.3 Préparation des données

Pour cette étape, une séance avec des facturistes et les codificatrices fut nécessaire, car les deux systèmes d'informations (Opale et Phoenix) contiennent de multiples variables pouvant être exploitées.

Voici une brève description des différents systèmes d'informations ci-dessus :

- **Le dossier administratif Opale SIAD**
  Dans le dossier administratif se trouvent toutes les données relatives à

l'administration du patient : sa date d'entrée et sa date de sortie de l'hôpital, sa durée de séjour, le montant qui est facturé au patient, son APDRG, etc. Lorsqu'il sera codé, c'est également dans ce système que se trouveront les codes diagnostiques et de traitements.

- **Le dossier clinique Phoenix SICL**
  Dans le dossier clinique se trouvent toutes les informations relatives aux traitements du patient : les médicaments prescris durant le séjour, les différentes prestations fournies au patient, ses analyses, etc.

A la suite de ces séances, un jeu de données potentiellement intéressantes fut déterminé (Tableau 6-1) :

Nom de la variable	Connu uniquement à la sortie du patient	Remarques
Le type d'admission		7 codes différents
Le mode d'entrée		24 codes différents
La provenance		84 codes différents
La décision d'envoi		8 codes différents
Le denre d'admission		9 codes différents
Le mode de sortie	X	35 codes différents
La destination	X	87 codes différents
La prise en charge après la sortie	X	9 codes différents
Le sexe		2 codes différents
La résidence pour convention		5 codes différents
L'âge à l'entrée		Calculé
Le type de cas		49 codes différents
La classe		8 codes différents
Le type de patient		31 codes différents
Le tarif		44 codes différents
Le type de taxe		39 codes différents
Le groupe de classe		3 codes différents
La division		35 codes différents
Le service		114 codes différents
L'unité		77 codes différents
Le médecin traitant		169 codes différents
La spécialité du médecin traitant		29 codes différents
Le genre de médecin traitant		12 codes différents
La durée de séjour		Calculé
La durée de séjour nette		Calculé
La liste des prestations Tarmed		4770 codes différents
Points de prestation Tarmed		Calculé
La valeur des prestations Tarmed		Calculé
Le Cost-Weight		Calculé
Le Cost-Weight pondéré		Calculé
Le total facturé		Calculé
Les heures de soins intensifs		Calculé
La lettre de sortie	X	Texte libre
Le diagnostic principal		Texte libre
Les comorbidités		Texte libre
L'intervention principale		Texte libre
L'intervention supplémentaire		Texte libre

Tableau 6-1 : Liste des variables disponibles pour la construction des modèles d'analyse

Mise à part les codes des services, des unités et des divisions qui répondent à une certaine hiérarchie, tous les autres codes sont indépendants les uns par rapport aux autres.

Après avoir contacté les personnes responsables des deux systèmes sources, celles-ci nous ont informés de la nature et de la saisie de certaines variables.

Par exemple, les données provenant du dossier clinique du patient (la lettre de sortie, le diagnostic principal, les comorbidités, l'Intervention principale, les interventions supplémentaires) sont des variables de texte libre, ce qui signifie qu'elles sont difficilement exploitables. De plus, elles ne sont pas nécessairement saisies dès l'admission du patient, durant le séjour ou immédiatement à la sortie, certaines ne sont simplement pas saisies dans le système.
Du fait de ces indications, nous avons décidé de ne pas utiliser les variables provenant du système clinique.

Les variables restantes possèdent, à la base, une meilleure qualité de saisie, car dans le système administratif, elles proviennent de champs comportant une liste de valeurs finie. De plus, la majorité des variables sont connues dès l'admission du patient, ce qui permet à tout moment d'évaluer les patients se trouvant dans les hôpitaux, même si ces variables peuvent changer au fil du séjour. Par contre, certaines informations du système administratif sont renseignées au fur et à mesure du séjour du patient (les prestations Tarmed).

Afin de ne pas perturber les systèmes sources, nous avons décidé d'utiliser le Data Warehouse qui comporte déjà la totalité des informations nécessaires à l'élaboration du modèle d'analyse (Figure 6-1). De plus, les données provenant des systèmes sources sont préalablement nettoyées avant d'y être introduites.

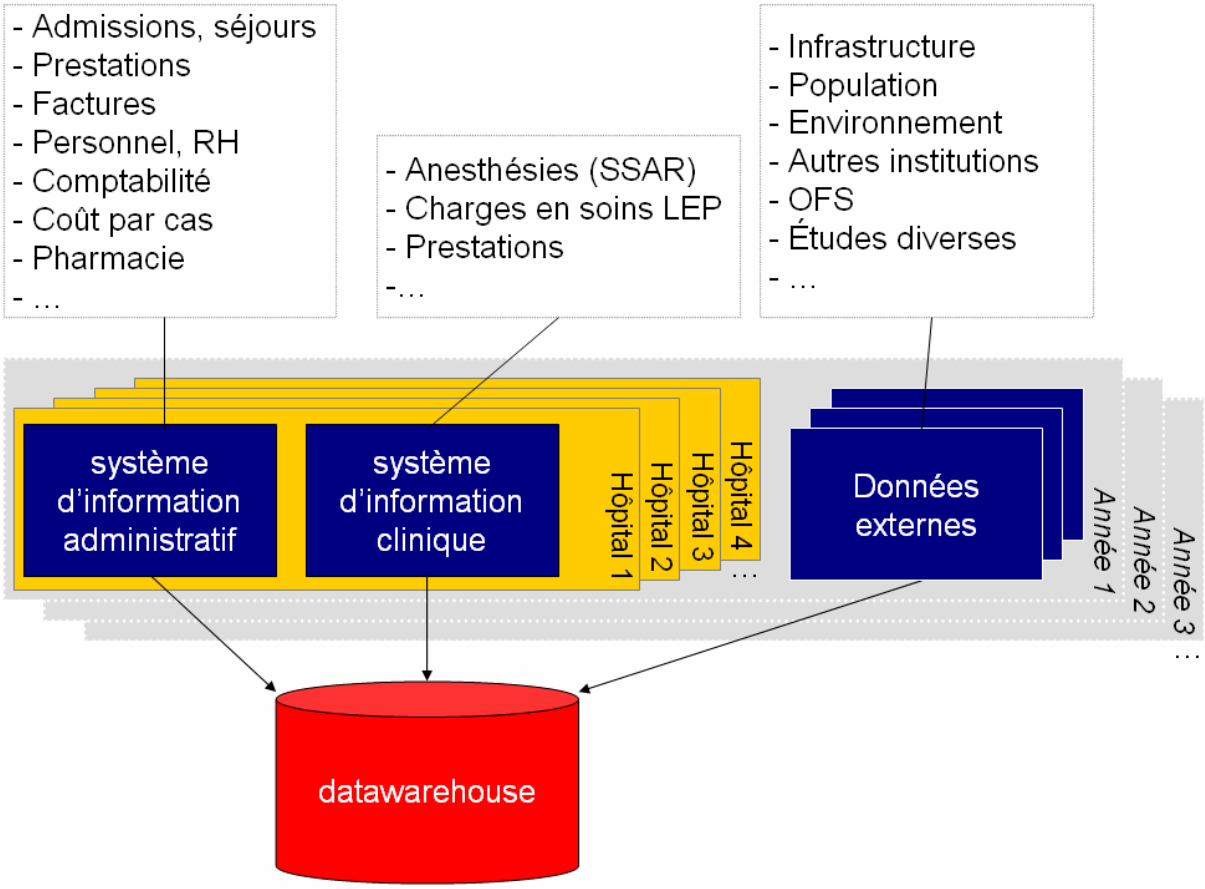


Figure 6-1 : Flux de chargement du Data Warehouse

Etant donné que notre système source est le Data Warehouse, nous le détaillons un peu en détail.
Le DW du SIMAV est en production depuis début 2003. Lors de sa mise en production, il contenait uniquement des données concernant les statistiques OFS (Office Fédérale de la Statistique) Médicale et Administrative des hôpitaux.
Ensuite, durant l'année 2003, la mise en production d'Opale et de Phoenix a eu lieu et c'est seulement à la fin de l'année que le DW commença à puiser des données à partir des deux nouveaux systèmes.
Début 2004, seules les données administratives des patients valaisans étaient extraites d'Opale. Mais durant la même année, les demandes d'informations contenues dans ce système se sont intensifiées et nous en avons extrait les données concernant les stocks, les ressources humaines et la comptabilité.
Parallèlement et toujours en 2004, nous avons extrait depuis Phoenix les informations nécessaires à l'élaboration des statistiques d'anesthésiologie. Durant l'année 2005, nous avons débuté l'extraction des données du LEP (Leistungserfassung in der Pflege – Saisie des prestations de soins infirmiers).

Il est intéressant de donner quelques caractéristiques des trois tables principales du DW qui sont utilisées pour ce projet :

- la table des séjours (> 788'000 enregistrements)
  Elle contient les informations comme : « la durée de séjour nette », « la durée de séjour », « L'age à l'entrée », « le sexe »
- la table des présences (> 3'067'000 enregistrements)
  Elle contient les informations comme : « Le cas », « la classe d'hospitalisation », « le type de patient », « le service ». On y trouve toutes les variables susceptibles de changer durant le séjour du patient.
- la table des prestations (> 38'025'000 enregistrements) :
  Elle contient toutes les prestations saisies durant le séjour d'un patient, comme : « Les prestation Tarmed », « Les analyses », « Les médicaments ».

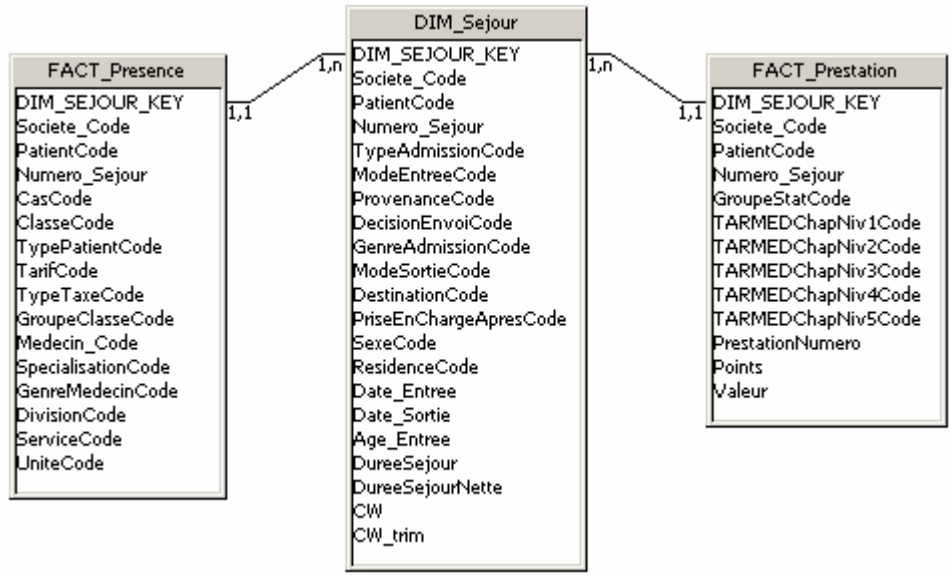


Figure 6-2 : Schéma de la base de données du Data Warehouse

Mis à part la table contenant les prestations qui est mis à jour hebdomadairement avec, en moyenne 210'000 enregistrements, les deux autres tables sont alimentées et rafraîchies quotidiennement : pour la table des séjours par > 1'000 enregistrements et pour la table des présences > 4'000 enregistrements.

Dans une optique de performance et aussi de constance dans les données de test, nous avons décidé de créer un Data Mart propre à l'utilisation de PrediRec. L'outil ETL pour le chargement des données est celui de la suite Microsoft SQL Server, c'est à dire SQL Server Integration Services (SSIS).

A l'aide de l'outil d'ETL, nous pouvons aussi préparer et stocker des données qui ne sont pas disponibles dans les systèmes sources, comme les heures de soins intensifs ou les durées de séjour des patients non sortis des hôpitaux.

Pour la préparation des données d'apprentissage, nous avons décidé d'effectuer une extraction par année de sortie (2006 et 2007) des cas déjà facturés et nous avons stocké ces informations dans des tables séparées (une table du Data Mart par année).

De ces données d'apprentissage, nous avons extrait aléatoirement 160 cas de tests pour l'année 2006 et 160 autres cas pour l'année 2007.
Les 160 cas sont répartis ainsi :

- 80 cas par centre hospitalier
  (En Valais, il existe, depuis début 2004, trois centres hospitaliers dont deux regroupent les hôpitaux de soins somatiques aigus : le SZO et le CHCVs. Ce sont uniquement ces deux derniers qui nous intéressent car seuls les cas somatiques aigus sont codés et facturés par APDRG.)
- parmi ces 80 cas, 40 sont considérés comme des Inliers
- des 40 cas restant, 20 sont considérés comme Low-Outliers
- les derniers 20 cas sont considérés comme des High-Outliers

La répartition par centre hospitalier permet, par exemple, d'évaluer si des différences de codage ou de lourdeur de cas peuvent être remarquées.

La répartition par Inliers/Outliers, permet de simuler le fait que les cas non codés à la fin d'une période sont à 50 % des cas Inliers et 50 % des cas Outliers. Le pourcentage d'Outliers a été fortement augmenté par rapport à la réalité, mais est justifié car ce sont ces cas qui doivent être estimés au plus proche en raison des montants prévisibles très élevés.

Le fait d'extraire ces cas directement des données d'apprentissage nous a permis d'utiliser, pour l'étape « 6.6 – Evaluation des différents modèle », des cas pour lesquels nous connaissions déjà la valeur des variables à estimer.

6.4 Création du modèle

Afin d'effectuer cette étape, nous devons ajouter une variable supplémentaire par rapport aux variables proposées dans le Tableau 6-1 : une clé primaire (unique) qui permet de différencier les cas les uns par rapport aux autres.

Après une recherche approfondie des algorithmes proposés par SSAS et l'exécution de divers tests, nous concluons que, parmi les algorithmes de prédiction, seuls les algorithmes « Microsoft Decision Trees », « Microsoft Times Series » et « Microsoft Neural Network » peuvent être utilisés pour la suite de ce projet.

Effectivement, seuls ces trois algorithmes permettent d'évaluer les colonnes continues, mais l'algorithme « Microsoft Times Series » ne peut être utilisé dans ce projet car nous ne cherchons pas à évaluer une tendance.

Nous commençons dès lors à créer un modèle d'exploration de données en définissant les types de données ainsi que les types de contenu des variables sélectionnées à l'étape précédente.

Pour préparer notre premier modèle, nous avons choisi de ne pas utiliser toutes les variables à notre disposition et nous les avons définies comme indiqué dans le Tableau 6-2 :

Nom de la variable	Type de données	Type de contenu	Utilisation
DIM_SEJOUR_KEY	Long	Key	Key
Le type d'admission	Text	Discrete	Input
Le mode d'entrée	Text	Discrete	Input
La provenance	Text	Discrete	Input
La décision d'envoi	Text	Discrete	Input
Le genre d'admission	Text	Discrete	Input
Le mode de sortie	Text	Discrete	Input
La destination	Text	Discrete	Input
La prise en charge après la sortie	Text	Discrete	Input
Le sexe	Text	Discrete	Input
La résidence pour convention	Text	Discrete	Input
L'âge à l'entrée	Long	Discretized	Input
Le type de cas	Text	Discrete	Input
La classe	Text	Discrete	Input
Le type de patient	Text	Discrete	Input
Le tarif	Text	Discrete	Input
La type de taxe	Text	Discrete	Input
Le groupe de classe	Text	Discrete	Input
La durée de séjour	Long	Continuous	Input
La durée de séjour nette	Long	Continuous	Input
Les heures de soins intensifs	Long	Continuous	Input
Le total facturé	Long	Continuous	PredictOnly
Le cost-Weight pondéré	Long	Continuous	PredictOnly

Tableau 6-2 : Variables utilisées pour la définition du premier modèle d'analyse

Nb. Le terme PredictOnly signifie que l'attribut en question est à prévoir et que celui-ci ne peut pas être utilisé en tant que variable d'entrée pour estimer une autre donnée.

Le premier algorithme testé fut celui de l'arbre de décision.

6.5 Exploration du modèle

Dans cette étape, nous avons parcouru l'onglet « Réseau de dépendance » pour comprendre comment l'algorithme « MDT » a exploité les variables fournies en entrée.

Une première observation fut le fait que l'algorithme nous a créé deux arborescences distinctes (Figure 6-3) : une pour prédire le « CW pondéré » et une autre pour estimer le « Total Facture », comme nous l'avions indiqué à l'étape précédente « 6.4 - Création du modèle ».

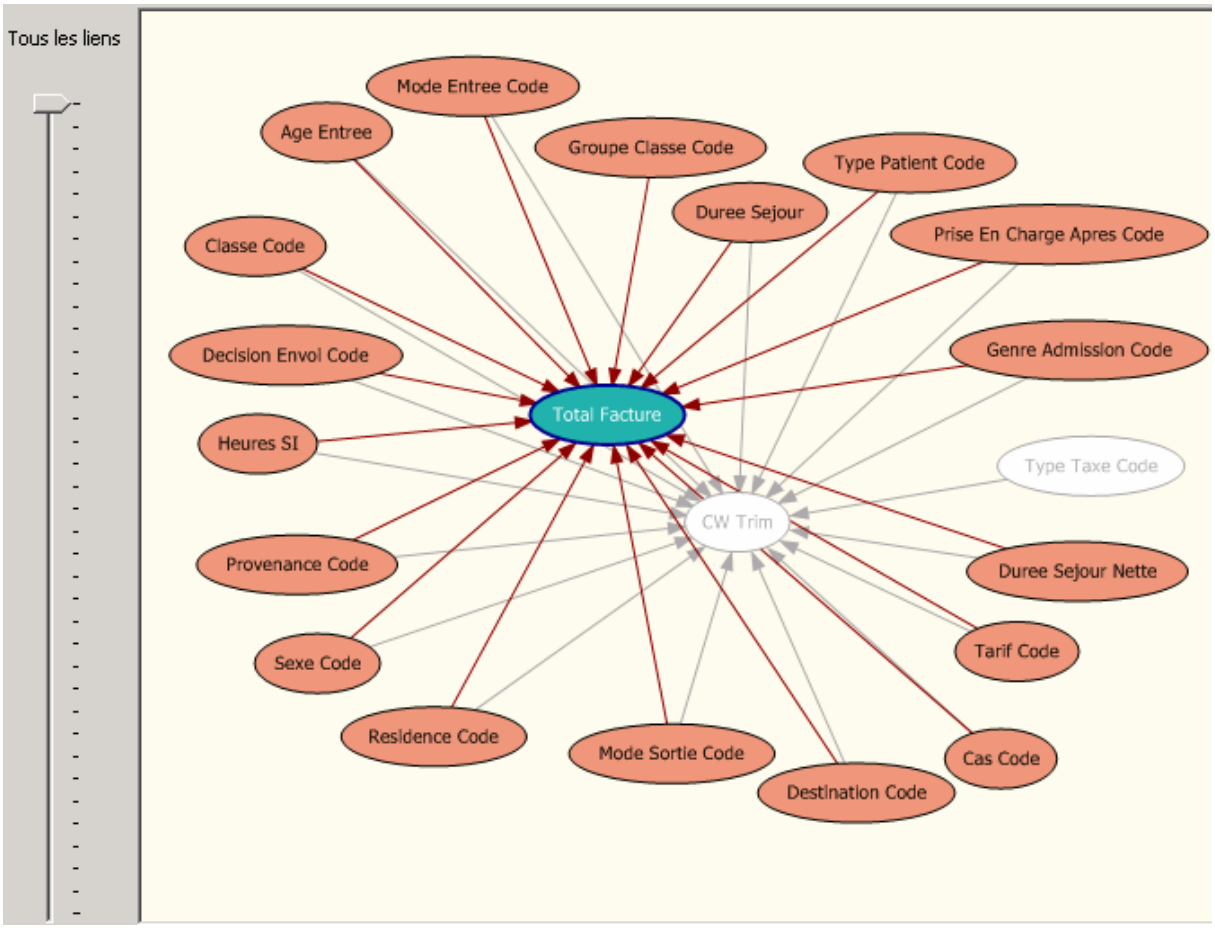


Figure 6-3 : deux arborescences distinctes

Nous remarquons aussi que certaines variables sont utilisées dans un modèle d'exploration de données et pas dans l'autre. Par exemple, « le type de taxe » n'est pas utilisé dans l'évaluation du « Total Facture » alors que, réciproquement, le « Groupe de classe » n'est pas utilisé dans l'estimation du « CW pondéré ».

A l'aide de la réglette à gauche de l'écran, nous constatons aussi que la « durée de séjour nette » correspond à la variable influençant le plus les deux modèles d'exploration de données (Figure 6-4).

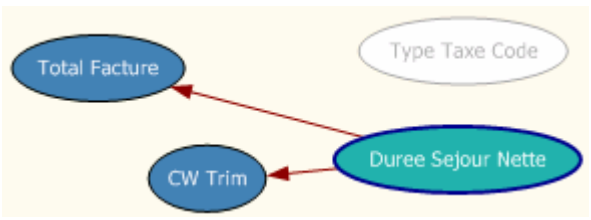


Figure 6-4 : Influence de la durée de séjour nette

Nous changeons d'onglet pour choisir l'« Arborescence de décision » et nous constatons que, par rapport au nœud affiché dans le tutorial, les nœuds ont changé d'aspect, passant de la Figure 6-5 à la Figure 6-6.

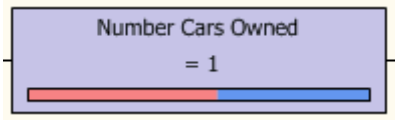


Figure 6-5 : Nœud affiché dans les tutoriaux

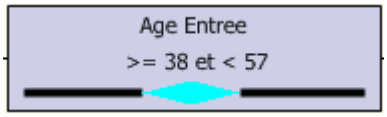


Figure 6-6 : Nœud affiché dans PrediRec

Après quelques recherches, nous trouvons une explication à cet affichage[15] :

- Figure 6-5 : La ligne représente un histogramme représentant la distribution des états de l'attribut prévisible.
  Cette figure représente la prédiction d'attributs discrets.
- Figure 6-6 : La ligne d'histogramme est remplacée par un losange. La largeur de celui-ci représente la variance de l'attribut du nœud et plus celui-ci est étroit, plus la prédiction des cas se trouvant dans ce nœud est précise.
  Cette figure représente la prédiction d'attributs continus.

[15] http://technet.microsoft.com/fr-fr/library/ms174503.aspx

6.6 Evaluation des différents modèles

Pour la phase de validation du modèle, nous avons préparé un classeur MS Excel contenant divers graphiques et autres informations. Nous sommes partis sur l'idée d'un classeur MS Excel car il nous facilite la tâche lors de la comparaison des divers modèles d'exploration de données par rapport à la visionneuse proposée par SSAS.

Ce fichier est composé de 5 onglets :

- un onglet contenant la fiche du test
- un onglet contenant les données
- un onglet contenant les graphiques liés à la prédiction de l'attribut « Total Facture »
- un onglet contenant les graphiques liés à la prédiction de l'attribut « CW pondéré »
- un onglet « feuille de travail »

Dans chacun de ces onglets nous trouvons un bouton « Imprimer » qui permet de masquer les onglets « Feuille de travail » et « Données » avant d'imprimer le test.

Nous détaillons ces différents onglets afin de comprendre sur quelles bases se sont déroulés les tests de validation.

Chaque test est sauvegardé sous une nomenclature précise. Le nom du fichier MS Excel commence par « PrediRec_Resultat_Test_ », est suivi du numéro du test et, si pour ce test nous avons essayé différents algorithmes de prédiction, l'abréviation de celui-ci.
Exemple : « PrediRec_Resultat_Test_024_02_MDT ».

6.6.1 Le classeur PrediRec_Resultat

L'onglet « Fiche signalétique »

La Figure 6-7 illustre l'onglet « Fiche signalétique ». Le rôle de cet onglet est de résumer quelles sont les variables qui ont été utilisées pour le test, quel est le nom de celui-ci, à quelle date il a eu lieu et quel a été l'algorithme de prédiction utilisé.

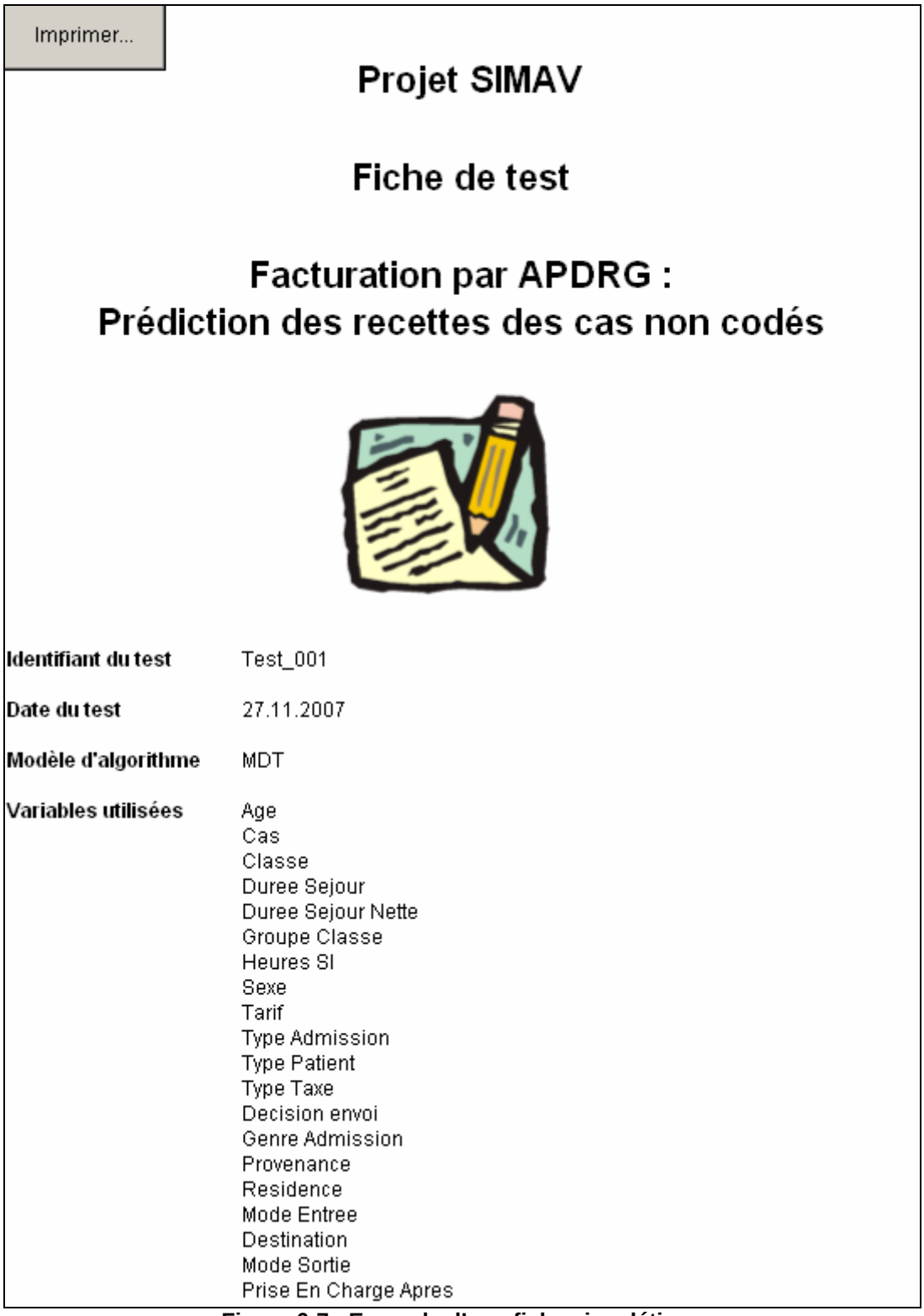


Figure 6-7 : Exemple d'une fiche signalétique

L'onglet « Données »

Cet onglet peut être partagé en deux parties (Figure 6-8) :

- La première partie, de la ligne 2 à la ligne 161, contient :
  - en vert clair, les valeurs des attributs que nous souhaitons approximer
  - en violet, les valeurs prédites par la méthode du RSV
  - en orange, les valeurs estimées par PrediRec

En complément des valeurs évaluées par les deux méthodes, nous avons complété l'onglet « Données » avec plusieurs calculs :

pour les deux méthodes de prévision :

- la différence entre l'estimation du « Total Facture » et sa valeur réelle
- la différence entre l'estimation du « CW pondéré » et sa valeur réelle
- 11 colonnes pour classer, en tranches de 100 Sfr., la différence en francs entre l'estimation du « Total Facturé » et la recette réelle
- 12 colonnes pour classer, en paliers de 20 % s'étalant de -100 % à + 100 %, les différences en % entre le « Total facture » et sa valeur réelle
- la différence en % du « Total facture » estimé par rapport à celui réel
- la différence en % du « CW Pondéré » estimé par rapport à celui réel

- La deuxième partie, de la ligne 165 à 176, contient plusieurs calculs permettant de comparer les méthodes RSV et PrediRec. Les calculs ci-dessous ont été exécutés une fois avec les valeurs relatives des différences et une autre fois avec les valeurs absolues de ces différences :
  - les centiles 10 / 25 / 50 / 75 / 90, le minimum, la moyenne et le maximum des différences (« Total Facture » estimées moins « Total Facture » réel) des « Total Facture » en francs
  - les centiles 10 / 25 / 50 / 75 / 90, le minimum, la moyenne et le maximum des différences (« CW Pondéré » estimées moins « CW Pondéré » réel) des « CW Pondéré »
  - les centiles 10 / 25 / 50 / 75 / 90, le minimum, la moyenne et le maximum des différences (« Total Facture » estimées moins « Total Facture » réel) absolues des « Total Facture » en francs
  - les centiles 10 / 25 / 50 / 75 / 90, le minimum, la moyenne et le maximum des différences (« CW Pondéré » estimées moins « CW Pondéré » réel) absolues des « CW Pondéré »
  - la différence en % de la somme des « Total Facture » estimées par rapport aux « Total Facture » réels
  - la différence en % de la somme des « CW Pondéré »estimées par rapport aux « CW Pondéré » réels

Des filtres automatiques sont affichés au sommet de cet onglet, permettant ainsi de filtrer des cas non codés selon divers critères (Centre hospitalier, cibler les différences inférieures à tant, etc.).

Dans la deuxième partie de la Figure 6-8, nous avons ajouté un bouton permettant d'exécuter une macro VBA qui prépare les données pour le calcul car, en cas d'utilisation de filtres MS Excel, ne sait pas mettre à jour automatiquement les centiles et tous les autres calculs.

SEJOUR_I	TotalFactur	CW_TrimD	RSV	CW_Trimi	PrediRec	CW_TrimN	TF-PrediRe	ABS(TI	Diff. %	CW-PrediRe		
938922	3'267.45	0.411	6'905.81	0.869	4'438.03	0.498	-1'170.58	1'170.58	35.83%	0.087	21.1%	7'950.00
941086	8'315.70	1.046	8'880.67	1.117	10'709.20	1.010	-2'393.50	2'393.50	28.78%	0.036	-3.5%	7'950.00
942635	2'869.95	0.361	7'151.99	0.900	2'888.30	0.328	-18.35	18.35	0.64%	0.033	-9.3%	7'950.00
943325	3'927.30	0.494	6'905.81	0.869	4'438.03	0.532	-510.73	510.73	13.00%	0.038	7.6%	7'950.00
	1'502'618.65	185.344	1'116'259.01	138.54	1'442'515.71	164.636						

	Différences				Différences ABS				Différence en %			
	TotalFacture	TotalFacture	CW	CW	TotalFacture	TotalFacture	CW	CW	TotalFacture	TotalFacture	CW	CW
Centile 10	- 4'415.18	- 4'024.41	- 0.540	- 0.397	303.27	246.84	0.038	0.029	-25.71%	-4.0%	-25.25%	-11.17%
Centile 25	- 2'563.09	- 1'517.00	- 0.326	- 0.137	1'221.09	584.74	0.150	0.064				
Centile 50	- 588.40	- 255.05	- 0.076	- 0.019	2'677.06	1'355.73	0.350	0.168				
Centile 75	2'941.50	1'031.86	0.357	0.200	5'144.65	3'579.44	0.613	0.398				
Centile 90	11'020.92	4'959.07	1.219	0.576	11'020.92	6'628.58	1.219	0.796				
Minimum	- 7'224.09	- 23'477.62	- 0.910	- 1.295	50.69	-	0.004	0.000				
Moyenne	2'414.75	375.64	0.293	0.129	5'452.77	3'187.74	0.678	0.371				
Maximum	98'799.69	68'113.50	12.421	8.033	98'799.69	68'113.50	12.421	8.033				

Calcul les percentiles selon sélection

Figure 6-8 : Exemple de l'onglet « Données »

Dans un souci de compréhension et d'aide à la navigation, les mêmes couleurs sont reprises dans les graphiques et les tableaux des onglets suivants.

L'onglet « Graphiques_TotalFacture »
Cet onglet réunit différents graphiques permettant de comparer visuellement la différence entre les méthodes RSV et PrediRec.

Au début de la première page, nous affichons, dans un ordre différent, les calculs de la deuxième partie de la Figure 6-8.

	Différences		Différences ABS	
	RSV	PrediRec	RSV	PrediRec
Minimum	- 7'224.09	- 23'477.62	50.69	-
Moyenne	2'414.75	375.64	5'452.77	3'187.74
Maximum	98'799.69	68'113.50	98'799.69	68'113.50
			1'116'259.01	1'442'515.71
Total facture		1'502'618.65	-34.61%	-4.17%

Figure 6-9 : Tableau récapitulatif des différences

	Différences		Différences ABS		
	RSV	PrediRec	RSV	PrediRec	
Centile 10	- 4'415.18	- 4'024.41	303.27	246.84	
Centile 25	- 2'563.09	- 1'517.00	1'221.09	584.74	Quartile 1
Centile 50	- 588.40	- 255.05	2'677.06	1'355.73	Quartile 2
Centile 75	2'941.50	1'031.86	5'144.65	3'579.44	Quartile 3
Centile 90	11'020.92	4'959.07	11'020.92	6'628.58	

Figure 6-10 : Tableau récapitulatif des centiles des différences

En dessous des tableaux récapitulatifs (Figure 6-9 et Figure 6-10), nous trouvons deux graphiques représentants les droites de régression des deux méthodes (Figure 6-11).

Ce graphique nous permet d'évaluer la précision des modèles utilisés. Un critère d'évaluation choisis est le R2 de la droite de régression.

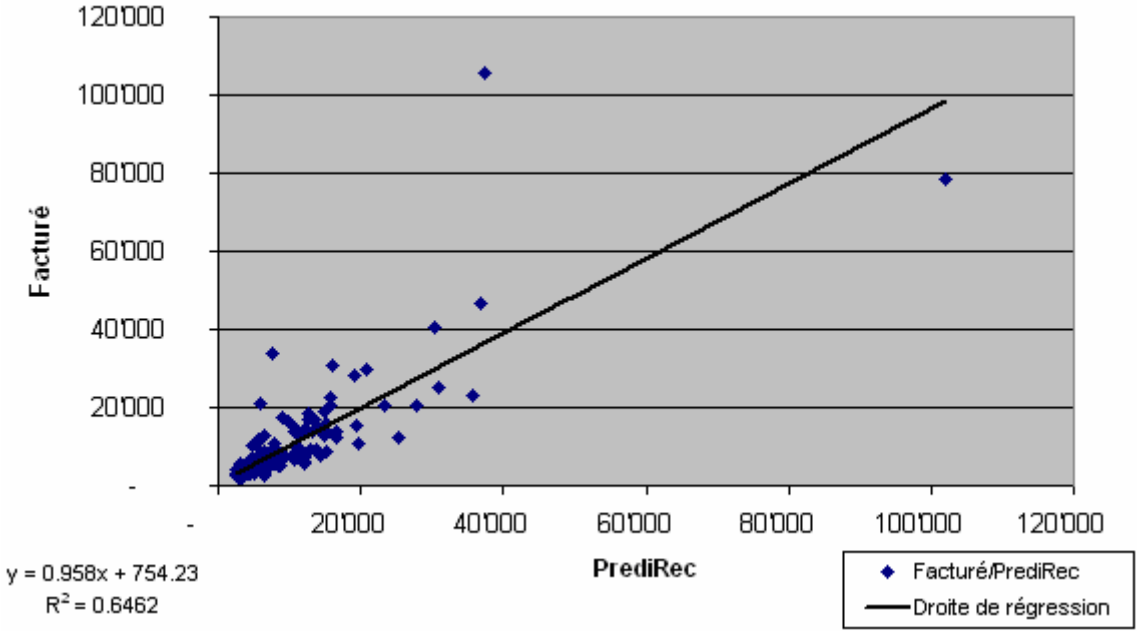


Figure 6-11 : Graphique de la droite de régression

A la suite du graphique de la droite de régression, nous affichons un autre graphique (Figure 6-12) comparant les résultats des méthodes RSV et PrediRec par rapport aux résultats attendus. De plus, ce graphique est trié dans l'ordre croissant des différences en % de la méthode PrediRec et la valeur réelle.

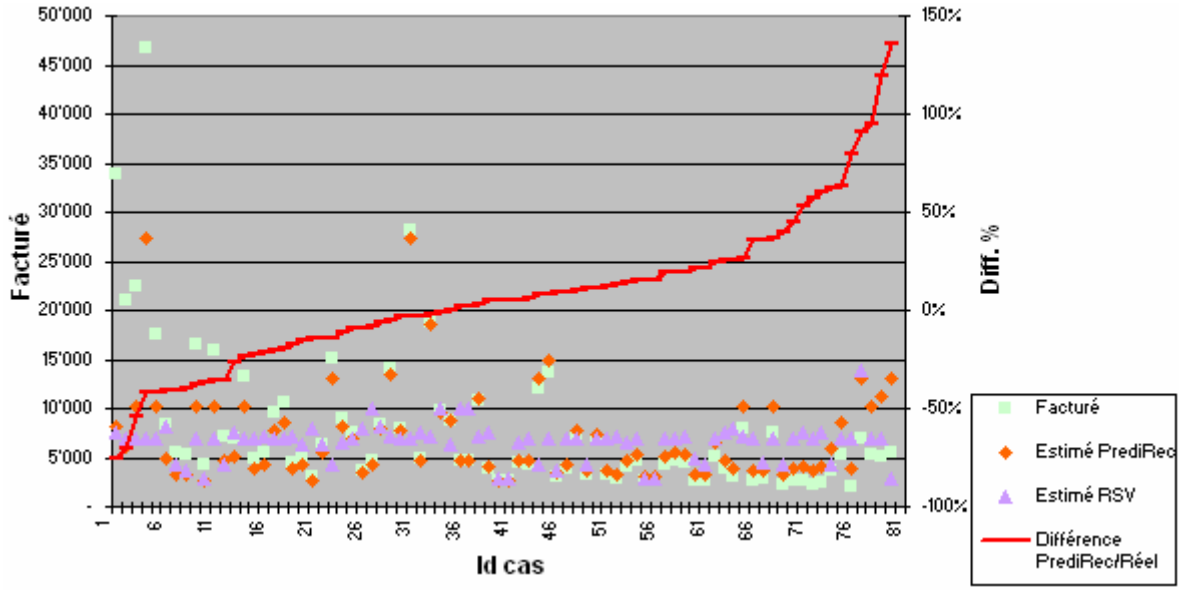


Figure 6-12 : Comparatif des estimations

Sur la page suivante, nous trouvons différents Box Plots représentant la dispersion des différences entre les valeurs estimées par les différentes méthodes par rapport aux valeurs escomptées.
Dans ce graphique, nous y affichons le centile 10 (le point plus à gauche du Box Plot), le centile 25 (la barre de gauche du rectangle) , le centile 50 (la barre rouge dans le rectangle, le centile 75 (la barre de droite du rectangle) et enfin le centile 90 (le point le plus à droite du Box Plot).

Les deux premiers Box Plots (Figure 6-13) représentent les différences relatives entre les valeurs prédites et les valeurs attendues.

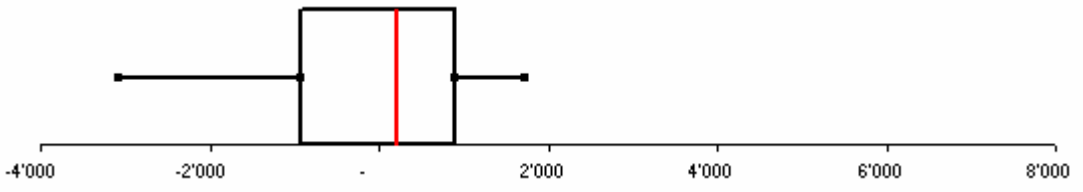


Figure 6-13 : Box Plot des différences

Les deux derniers Box Plots (Figure 6-14) représentent les différences absolues entre les valeurs estimées et les valeurs réelles.

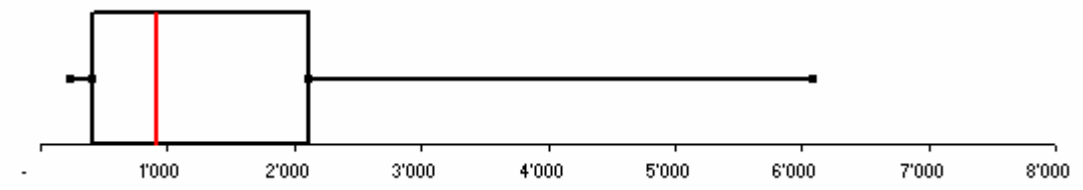


Figure 6-14: Box Plot des différences absolues

A la suite des Box Plots, nous trouvons le graphique suivant (Figure 6-15) qui groupe les cas non codés selon leur différence absolue en diverses tranches de 100 Sfr. :

- moins que 100 (<100) ;
- entre 100 et 499 (100 - 499) ;
- entre 500 et 999 (500 - 999) ;

- entre 1000 et 1999 (1000 - 1999) ;
- entre 2000 et 2999 (2000 - 2999) ;
- entre 3000 et 3999 (3000 - 3999) ;
- entre 4000 et 4999 (4000 - 4999) ;
- entre 5000 et 9999 (5000 - 9999) ;
- entre 10000 et 14999 (10000 - 14999) ;
- entre 15000 et 19999 (15000 - 19999) ;
- plus de 20000 (> 20000).

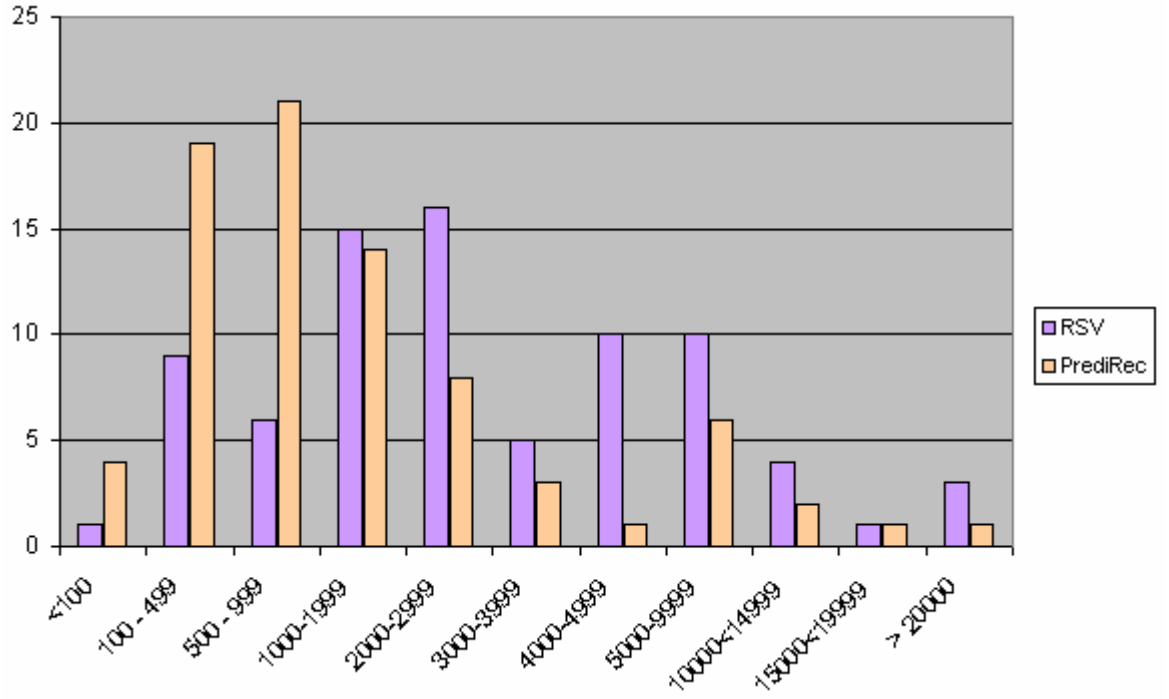


Figure 6-15 :Différence absolue par tranche

Le dernier graphique (Figure 6-16) affiche une courbe représentant les différences relatives groupées en classes de 20 % :

- plus de 100 % de différences sous estimées (-100 %) ;
- entre 100 % et 80 % de différences sous estimées (-100 - 80 %) ;
- entre 80 % et 60 % de différences sous estimées (-80 - 60 %) ;
- entre 60 % et 40 % de différences sous estimées (-60 - 40 %) ;
- entre 40 % et 20 % de différences sous estimées (-40 - 20 %) ;
- entre 20 % et 00 % de différences sous estimées (-20 - 00 %) ;
- entre 00 % et 20 % de différences sur estimées (00 - 20 %) ;
- entre 20 % et 40 % de différences sur estimées (20 - 40 %) ;
- entre 40 % et 60 % de différences sur estimées (40 - 60 %) ;
- entre 60 % et 80 % de différences sur estimées (60 - 80 %) ;
- entre 80 % et 100 % de différences sur estimées (80 - 100 %) ;
- plus de 100 de différences sur estimées (+100 %).

Ce graphique permet de vérifier que les résultats du modèle d'analyse sont bien distribués de manière normale (selon un courbe de Gauss) autour d'un mode se situant le plus proche possible de 0.

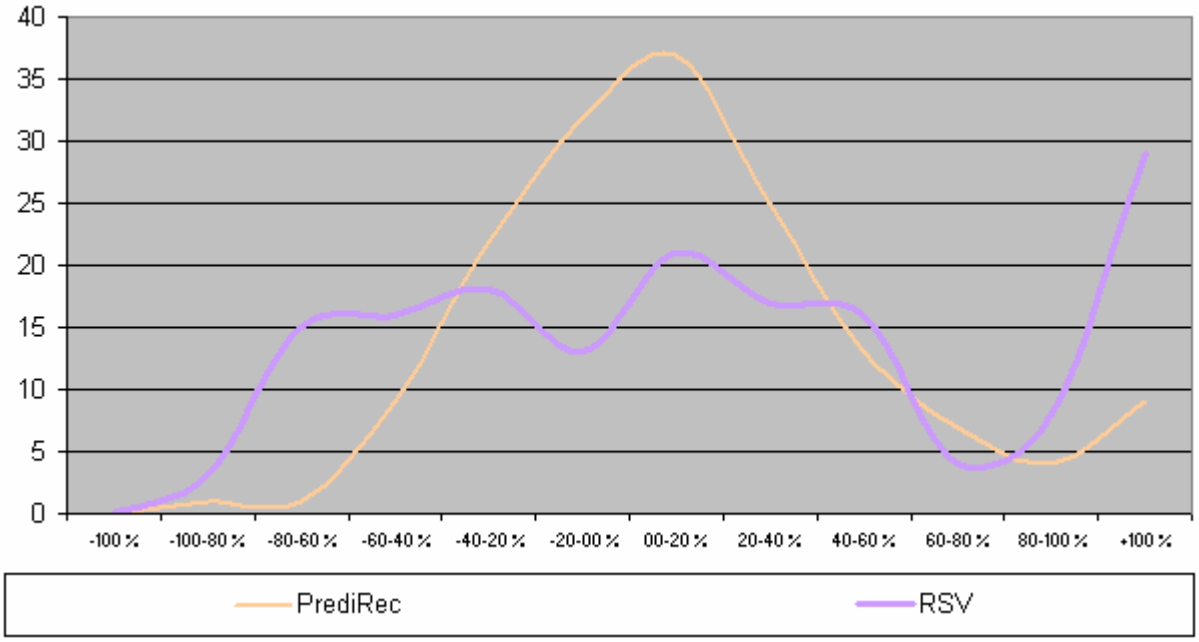


Figure 6-16 : Différence par classe de 20 %

L'onglet « Graphiques_CW »
Cet onglet reprend et affiche tous les graphiques de l'onglet « Graphiques_TotalFacture », sauf celui de la Figure 6-15 « Différence ABS par classe ».

Tous les graphiques de l'onglet représentent les « CW pondéré » en lieu et place des « Total Facture ».

L'onglet « FeuilleTravail »
Cet onglet nous sert à trier les cas non codés selon leur clé avant de les reprendre dans l'onglet « Données ».

Un exemple de fichier MS Excel se trouve sur le CD-ROM fourni avec ce rapport de projet.
Il est situé dans le dossier « Documents annexes » et se nomme « PrediRec_Resultat_Test_024_02_MDT.xls »

6.6.2 Les tests des modèles d'analyses

Avant de commencer différents tests de modèles d'analyses, nous avons commencé par grouper les différentes variables du Tableau 6-1 en catégories, selon leur utilité dans les systèmes sources.

Le Tableau 6-3 résume les différentes catégories choisies ainsi que les variables qui s'y réfèrent :

Catégorie	Liste des variables
Données administratives (OPALE)	Age, Cas, Classe, Duree Sejour, Duree Sejour Nette, Groupe Classe, Heures SI, Sexe, Tarif, Type Admission, Type Patient, Type Taxe
Variables d'entrées (OPALE)	Decision envoi, Genre Admission, Provenance, Residence, Mode Entree
Variables de sorties (OPALE)	Destination, Mode Sortie, Prise En Charge Apres
Prestations Tarmed (OPALE)	Liste des prestations Tarmed Points de prestation Tarmed Valeur des prestations Tarmed
Informations PHOENIX	Lettre de sortie, Diagnostic principal, Comorbidités, Intervention principale, Intervention supplémentaire
Données sur le médecin (OPALE)	Medecin, Specialisation, Genre
Variables sur le service (OPALE)	Service, Unité

Tableau 6-3 : Catégorie des variables

Comme expliqué au chapitre « 6.3 - Préparation des données », nous n'avons pas utilisé les variables de la catégorie « Informations PHOENIX », provenant du système source Phoenix, car ces diverses données ne sont pas forcément disponibles lors de l'évaluation d'un cas.

Pour ce projet, nous avions deux modèles d'analyse à valider :

- un pour le « Total Facture » ;
- un pour le « CW pondéré ».

Nous sommes aussi parti du principe que, selon le modèle d'analyse, les variables à utiliser ne sont pas les mêmes, comme nous l'avons constaté au chapitre « 6.5 - Exploration du modèle ».

Afin de présenter plusieurs estimations au chef de la division « Finance et Controlling » du SZO ainsi qu'au chef de service du SIMAV, nous avons effectué plus d'une quarantaine de tests pour lesquels nous avons repris les résultats dans le classeur MS Excel présenté au chapitre précédent.

Voici la liste des différents tests effectués :

Nom	Algorithmes	Variables utilisées
Test_01	MDT	Données administratives + VariablesSorties + VariablesEntrées
Test_02	MDT	Données administratives + VariablesEntrées
Test_03	MDT	Données administratives
Test_04_01	MDT	Données administratives + VariablesSorties + VariablesEntrées + Medecin + Specialisation + Genre
Test_04_02	MDT	Données administratives + VariablesSorties + VariablesEntrées + Medecin + Specialisation
Test_04_03	MDT	Données administratives + VariablesSorties + VariablesEntrées + Medecin
Test_04_04	MDT	Données administratives + VariablesSorties + VariablesEntrées + Specialisation
Test_05_01	MDT	Données administratives + VariablesEntrées + Medecin + Specialisation + Genre
Test_05_02	MDT	Données administratives + VariablesEntrées + Medecin + Specialisation
Test_05_03	MDT	Données administratives + VariablesEntrées + Medecin
Test_05_04	MDT	Données administratives + VariablesEntrées + Specialisation
Test_06_01	MDT	Données administratives + Medecin + Specialisation + Genre
Test_06_02	MDT	Données administratives + Medecin + Specialisation
Test_06_03	MDT	Données administratives + Medecin
Test_06_04	MDT	Données administratives + Specialisation
Test_07_01	MDT	Données administratives + VariableSorties + VariablesEntrées + ServiceSortie + UniteSortie
Test_07_02	MDT	Données administratives + VariableSorties + VariablesEntrées + ServiceSortie
Test_07_03	MDT	Données administratives + VariableSorties + VariablesEntrées + UniteSortie
Test_08_01	MDT	Données administratives + VariablesEntrées + ServiceSortie + UniteSortie
Test_08_02	MDT	Données administratives + VariablesEntrées + ServiceSortie
Test_08_03	MDT	Données administratives + VariablesEntrées + UniteSortie
Test_09_01	MDT	Données administratives + ServiceSortie + UniteSortie
Test_09_02	MDT	Données administratives + ServiceSortie
Test_09_03	MDT	Données administratives + UniteSortie
Test_010	MDT	Données administratives + GroupeStatistiqueTarmed
Test_011	MDT	Données administratives + Niveau1Tarmed
Test_012	MDT	Données administratives + Niveau2Tarmed
Test_013	MDT	Données administratives + Niveau3Tarmed
Test_014	MDT	Données administratives + Niveau4Tarmed
Test_015	MDT	Données administratives + Niveau5Tarmed
Test_023	MDT-MNN	Données administratives
Test_024_04	MDT-MNN	Données administratives + VariablesSorties + VariablesEntrées + Specialisation
Test_025_04	MDT-MNN	Données administratives + VariablesEntrées + Specialisation
Test_026_04	MDT-MNN	Données administratives + Specialisation
Test_030	MLR-MNN	Données administratives + Specialisation

Données administratives :
Age, Cas, Classe, Duree Sejour, Duree Sejour Nette, Groupe Classe, Heures SI, Sexe, Tarif, Type Admission, Type Patient, Type Taxe

VariablesSorties
Destination, Mode Sortie, Prise En Charge Apres

VariablesEntrées
Decision envoi, Genre Admission, Provenance, Residence, Mode Entree

Le déroulement des tests se passa ainsi.

Les trente premiers tests furent effectués uniquement avec l'algorithme d'arbre de décision de Microsoft.

Test_001
Nous avons commencé par tester les variables administratives ainsi que les variables d'entrées et les variables de sorties.

Test_002
Après le premier test, nous avons réfléchi à la problématique des cas non codés qui ne sont pas encore sortis de l'hôpital lors de l'évaluation. Suite à cette réflexion, nous avons supprimé les variables de sorties, ce qui représente le deuxième test.

Test_003
A la suite du deuxième test, nous voulions aussi tester un modèle d'analyse sans les variables d'entrées, pas forcément pertinentes, ce qui entraîna le troisième test.

Première analyse
Après avoir repris les différents résultats dans le classeur MS Excel, nous les avons comparés. Entre le premier et le deuxième test, les résultats pour le « Total Facture » et le « CW pondéré » étaient totalement égaux. Pour comprendre cette égalité, nous avons analysé le réseau de dépendance (Figure 6-3) du premier test et nous avons constaté que les variables de sorties n'étaient pas utilisées dans ce modèle d'analyse.

Des trois premiers tests, le troisième était le plus précis, mais il n'est pas encore satisfaisant, car l'un des critères d'évaluation était la volonté d'avoir un R2 (dans le graphique de la droite de régression Figure 6-11) de 0.80 au moins, alors qu'il n'était que de 0.6505 pour le « Total Facture » et de 0.7354 pour le « CW pondéré ».

Test_004_01
Afin d'améliorer les résultats, nous avons repris le premier test et nous avons ajouté les données des médecins.

Test_004_02
De ce nouveau modèle d'analyse, nous avons éliminé la variable « Genre », car celle-ci était utilisée tardivement par l'algorithme.

Test_004_03 + Test_004_04
Après avoir analysé le réseau de dépendance, nous avons remarqué que les variables « Spécialisation » et « Médecin » interviennent rapidement dans le modèle d'analyse. Nous avons donc créé, pour chacune des deux variables, un modèle d'analyse propre.

Deuxième analyse
Nous avons comparé les différents résultats obtenus mais nous n'arrivions toujours pas à notre objectif du R2 supérieur à 0.80. Nos meilleurs résultats furent 0.7873 pour le « Total Facture » (Test_004_04) et 0.7436 pour le « CW pondéré » (Test_004_01).

Test_005_01 + Test_005_02 + Test_005_03 + Test_005_04
Etant donné que les tests « Test_004_0x » sont construits sur la base du premier test et que nous devons aussi estimer des cas non sortis des hôpitaux, nous avons adapté les « Test_004_0x » pour créer les « test_005_0x » en supprimant les variables de sorties.
Dans un souci de compréhension, nous avons construit les « Test_005_0x » avec la même logique appliquée pour les « Test_004_0x » correspondant.

Troisième analyse
Nous avons à nouveau comparé les différents résultats obtenus mais nous n'arrivions toujours pas à notre objectif. Nos meilleurs résultats furent 0.7832 pour le « Total Facture » (Test_005_04) et 0.7815 pour le « CW pondéré » (Test_005_04).
Nous avons constaté que les modèles d'analyse du « Total Facturé » étaient moins précis que celui du « Test_004_04 ».

Test_006_01 + Test_006_02 + Test_006_03 + Test_006_04
Puisque les tests « Test_004_0x » et « Test_005_0x » sont basés respectivement sur le premier et le deuxième test et que le troisième test fut le plus précis, nous avons construit les tests « Test_006_0x » sur la base de ce dernier.
Comme pour les « Test_005_0x », nous avons repris la même logique que celle appliquée pour les « Test_004_0x »

Quatrième analyse
Nous avons à nouveau comparé les différents résultats obtenus mais nous n'arrivions toujours pas à notre objectif. Nos meilleurs résultats furent 0.7867 pour le « Total Facture » (Test_006_04) et 0.7483 pour le « CW pondéré » (Test_006_04).

A ce stade, nous avons aperçu que les modèles d'analyse du « Total Facturé » étaient toujours moins précis que celui du « Test_004_04 » et que le meilleur « CW pondéré » est toujours celui du test « Test_005_04 »

Test_007_01 + Test_007_02 + Test_007_03
Test_008_01 + Test_008_02 + Test_008_03
Test_009_01 + Test_009_02 + Test_009_03
Lors de l'entretien avec les facturistes, ceux-ci ont soulevé le fait que le service ou l'unité d'où le patient est sorti de l'hôpital peut être un bon indicateur. Nous avons donc, comme pour les variables des médecins, repris les trois premiers tests en ajoutant les variables « Service » et « Unité »

Cinquième analyse
Après avoir créé les neuf tests ci-dessus, nous avons observé les résultats et nous avons constaté qu'aucun des modèles utilisant le service ou l'unité n'était concluant. Tous les R2 de ces modèles d'analyses étaient compris entre 0.64 et 0.66 pour le « Total Facture » et compris entre 0.74 et 077 pour le « CW pondéré ».

Les tests suivants, du « Test_010 » au « Test_020 » sont basés uniquement sur les variables administratives (« Test_003 ») et le cumul des points des prestations Tarmed selon des groupements différents.

Test_010
Dans ce test, nous avons cumulé les points des prestations Tarmed selon le « groupe statistique Tarmed ».

Test_011
Dans ce test, nous avons cumulé les points des prestations Tarmed selon le premier niveau de groupement des prestations Tarmed (niveau 1 Tarmed).

Test_012
Dans ce test, nous avons cumulé les points des prestations Tarmed selon le deuxième niveau de groupement des prestations Tarmed.

Test_013
Dans ce test, nous avons cumulé les points des prestations Tarmed selon le troisième niveau de groupement des prestations Tarmed.

Test_014
Dans ce test, nous avons cumulé les points des prestations Tarmed selon le quatrième niveau de groupement des prestations Tarmed.

Test_015
Dans ce test, nous avons cumulé les points des prestations Tarmed selon le cinquième niveau de groupement des prestations Tarmed.

Sixième analyse
Nous pensions que le fait d'inclure les points des prestations Tarmed augmenterait la précision des modèles d'analyses. Mis à part pour le « Test_010 », qui fut notre premier test à passer le cap fatidique d'un R2 supérieur à 0.80 (il était de 0.8023) pour le « Total Facture », les autres tests furent décevants et même catastrophiques. Par exemple, les R2 du « Total Facture » dégringole de 0.7624 pour « Test_011 » à 0.4156 pour « Test_015 ».
Par contre, le « CW pondéré » des tests « Test_010 » et « Test_011 » furent tous deux supérieurs à un R2 de 0.80, respectivement 0.8084 et 0.8503.
Malheureusement, après avoir demandé des précisions sur la saisie des prestations Tarmed, nous dûmes les laisser tomber, car celles-ci ne sont pas saisies au fur et à mesure du séjour. Elles peuvent même être saisies, dans certains cas, après la facturation du séjour au patient.

N'étant toujours pas satisfait de ces résultats, en dehors de ceux obtenus par les tests « Test_010 » et « Test_011 », nous avons décidé de changer de type de contenu pour les variables de types continues (« Age », « Durée Séjour », « Durée Séjour Nette » et « Heures de soins intensifs »). Nous les avons passées de « Continuous » à « Discretized ». Dans SSAS, il existe une méthode qui permet de discrétiser des valeurs continues en valeurs discrètes. Elle consiste à classer des valeurs continues dans des classes afin d'obtenir un nombre fini (discret) d'états possibles[16].
()

Afin de ne pas recommencer tous les tests, nous avons repris ceux qui avaient obtenus les meilleurs résultats précédemment, c'est à dire :

- Test_003 ;
- Test_004_04 ;
- Test_005_04 ;
- Test_006_04.

Ces différents tests sont donc adaptés afin d'utiliser la méthode de discrétisation pour les valeurs continues. Nous avons profité de ces nouveaux tests pour d'essayer l'algorithme neuronal de Microsoft.

Chacun des tests originaux a donc donné naissance à deux autres tests et, afin de reconnaître de quels tests sont issus les nouveaux, nous avons simplement ajouté 20 à l'identifiant du test :

- Test_003 Test_023_MDT + Test_023_MNN ;
- Test_004_04 Test_024_04_MDT + Test_024_04_MNN ;
- Test_005_04 Test_025_04_MDT + Test_025_04_MNN ;
- Test_006_04 Test_026_04_MDT + Test_026_04_MNN.

[16] http://technet.microsoft.com/fr-fr/library/ms174512.aspx

Septième analyse
Après avoir exécuté les derniers tests, nous les avons comparés, toujours pour trouver le meilleur modèle d'analyse. Tous les tests ont obtenu, pour le « Total Facture » des R2 supérieurs à 0.80 avec l'algorithme d'arbre de décision de Microsoft. Tous les résultats obtenus avec l'algorithme de réseau neuronal de Microsoft sont en dessous de la barre des 0.80. Le Tableau 6-4 résume les résultats obtenus :

Test	R2 obtenu
Test_023_MDT	0.8368
Test_023_MNN	0.7827
Test_024_04_MDT	0.8572
Test_024_04_MNN	0.7515
Test_025_04_MDT	0.8611
Test_025_04_MNN	0.7676
Test_026_04_MDT	0.8437
Test_026_04_MNN	0.7668

Tableau 6-4 : Résultats des modèles testés avec la méthode de discrétisation

Par contre, seul le « CW pondéré » du test « Test_023_MDT » franchit la barre du R2 supérieur à 0.80.

Une fois tous les tests effectués, nous avons synthétisé les résultats de tous les tests dans un fichier MS Excel « Synthèse des test.xls ». Vous pouvez trouvez ce fichier sur le CD-ROM, dans le dossier « Documents annexes ».
A partir de ce fichier, nous avons sélectionné les quatre meilleurs résultats du R2 pour le « Total Facture » ainsi que les quatre meilleurs pour le « CW pondéré » et nous les avons présentés au chef de service du SIMAV pour la validation des choix des modèles d'analyse finaux. La validation effectuée de la part du responsable administratif, finance et qualité du SZO sera faite durant le premier trimestre 2008.

Voici la liste des meilleurs résultats pour l'estimation du « Total Facture » :

Test	R2 obtenu
Test_025_04_MDT	0.8611
Test_024_04_MDT	0.8572
Test_026_04_MDT	0.8437
Test_023_MDT	0.8368

Tableau 6-5

Voici la liste des meilleurs résultats pour l'estimation du « CW pondéré » :

Test	R2 obtenu
Test_023_MDT	0.8073
Test_005_04	0.7815
Test_025_04_MDT	0.7787
Test_008_01	0.7755

Tableau 6-6

Durant la séance de validation, nous avons ajouté un deuxième critère pour le choix des modèles d'analyses :

- La différence relative totale, en %, soit entre le « Total Facture » ou « CW pondéré », par rapport aux résultats attendus.

Test	Total Facture Réel	Total Facture Estimé	Différence Sfr.	Différence %
Test_025_04_MDT	1'502'628.65	1'440'071.73	- 62'546.92	- 4.2 %
Test_024_04_MDT	1'502'628.65	1'447'251.70	- 55'366.95	- 3.7 %
Test_026_04_MDT	1'502'628.65	1'462'230.01	- 40'388.64	- 2.7 %
Test_023_MDT	1'502'628.65	1'471'004.33	- 31'614.32	- 2.1 %

Tableau 6-7 : Résultats des quatre meilleurs modèles d'analyse pour l'estimation du « Total Facture »

Test	CW Pondéré Réel	CW Pondéré Estimé	Différence Point	Différence %
Test_023_MDT	185.34	174.481	- 10.863	- 5.9 %
Test_005_04	185.34	167.182	- 18.162	- 9.8 %
Test_025_04_MDT	185.34	169.598	- 15.746	- 8.5 %
Test_008_01	185.34	167.796	- 17.548	- 9.5 %

Tableau 6-8: Résultats des quatre meilleurs modèles d'analyse pour l'estimation du « CW pondéré »

Le Tableau 6-7 résume les résultats des quatre meilleurs tests pour la prédiction du « Total Facture ».

Le Tableau 6-8 reprend, quant à lui, les quatre meilleurs tests pour l'estimation du « CW pondéré ».

6.7 Déploiement et mise à jour du modèle

Une difficulté de ce projet fut le déploiement des modèles d'analyses.

Etant donné que la solution mise à disposition des utilisateurs est un site Internet, il a été nécessaire de traduire les deux modèles générés par Visual Studio Business Intelligence en langage DMX (Data Mining eXtension).
Cette transformation fut nécessaire, car nous avions définis qu'un utilisateur possède ses propres modèles d'analyses afin qu'il puisse, pendant plusieurs jours, simuler des cas non codés et arriver toujours sur les mêmes résultats.
Si nous étions partis de l'idée que chacun des utilisateurs utilise un modèle d'analyse commun et que nous le mettions à jour avec les nouvelles données disponibles quotidiennement depuis le système source, la simulation des cas non codés serait différente d'un jour à l'autre.

Il en découle donc que l'utilisateur de PrediRec est responsable de mettre à jour lui-même ces modèles d'analyses. De plus, la mise à jour des modèles passe quasi inaperçue, car le site Internet intègre les fonctionnalités de chargements et de traitements nécessaires et l'utilisateur les utilise sans s'en rendre compte.

Pour ce qui est des temps d'exécution pour effectuer les mise à jour des modèles d'analyses, ceux –ci sont tout à fait correct : il faut compter moins d'une minute pour :

- extraire depuis le Data Warehouse les données nécessaires à l'apprentissage (> 27'000 cas codés)
- procéder à la mise à jour des modèles d'analyses

Ensuite, l'estimation d'environ +/- 2000 cas non codés dure :

- environs dix secondes pour extraire les cas non codés du Data Warehouse
- deux secondes pour présenter ces cas à PrediRec et afficher les résultats sur le site Web

Durant cette étape, nous avons dû définir quels sont les cas qui doivent être extrait du Data Warehouse, soit pour la phase d'apprentissage, soit pour la phase de prédiction. Les cas (séjours) retenus sont :

- les hospitalisations
- les naissances
- les types de cas somatiques aigus :
    - 0299 Cardiologie n/cantonalisée ;
    - 0300 Cardiologie cantonalisée ;
    - 0301 Chirurgie cardiaque cantonalisée ;
    - 0302 Chirurgie du dos (Neurochirurge) ;
    - 0303 Chirurgie Générale ;
    - 0304 Chirurgie maxillo-faciale ;
    - 0306 Chirurgie pédiatrique ;
    - 0307 Chirurgie Plastique/Reconstructive ;
    - 0308 Chirurgie thoracique ;
    - 0309 Chirurgie vasculaire ;
    - 0310 Chirurgie Esthétique ;
    - 0311 Chirurgie cardiaque n/cantonalisée ;
    - 0402 Dentaire ;
    - 0700 Gastro-entérologie ;
    - 0710 Gynécologie ;
    - 1201 Lithotripsie ;

  - 1300 Médecine ;
  - 1302 Médecine / Oncologie ;
  - 1400 Néonatologie ;
  - 1401 Neurochg. Spécialisée ;
  - 1402 Nurserie ;
  - 1405 Neurochg. Générale ;
  - 1406 Neurologie ;
  - 1500 Maternité-Obst. ;
  - 1501 Oncologie Lourde ;
  - 1502 Ophtalmologie ;
  - 1503 ORL ;
  - 1504 Orthopédie ;
  - 1600 Pédiatrie ;
  - 1603 Pneumologie ;
  - 1805 Radio-oncologie ;
  - 2000 Traumatologie orthopédique ;
  - 2100 Urologie.
- tous les types de patients non facturés par APDRG, soit différents de :
  - 42 Etranger / UE Sans E111 ;
  - CE Convention Esthétique.
- prendre les variables disponibles lors de la sortie
  Si le patient n'est pas sorti, calcul automatique de la durée de séjour, de la durée de séjour nette selon la date du jour de l'extraction des données, ainsi que les variables disponibles dans le système ce jour-là.

Si un cas (patient) répond aux exigences ci-dessus, celui-ci rentre dans le cadre d'une facturation par APDRG et peut en conséquence être estimé ou utiliser pour l'apprentissage.

7 Choix des modèles finaux

Les modèles d'analyses choisis pour PrediRec sont :

- pour le « Total Facture » : Test_023_MDT ;
- pour le « CW pondéré » : Test_023_MDT.

Argumentation des choix

Nous avons sélectionné le modèle d'analyse « Test_023_MDT » pour l'estimation du « Total Facture » car celui-ci possède un R2 supérieur à 0.80 (0.8368), mais, même s'il n'affiche pas le meilleur R2, il a l'avantage d'estimer le total relatif du « Total Facture » à seulement - 2.71 % des résultats attendus.

Nous avons choisi le modèle d'analyse « Test_023_MDT » pour la simulation du « CW pondéré » car il possède le meilleur R2 (0.8073) et la différence relative entre le « CW pondéré » estimé (174.481) et le « CW pondéré » réel (185.34) n'est que de -5.9 % (-10.863).

Il est important de souligner que le fait que ce soit le même modèle d'analyse choisi pour l'estimation du « Total Facture » et du « CW pondéré » n'est pas une simplification en soi étant donné que nous avons de toute façon deux variables à estimer et qu'il est nécessaire d'implémenter deux modèles d'analyses dans SSAS.

Les variables fournies en entrées sont les suivantes, organisée selon leur ordre d'importance dans les modèles d'analyses :

- La durée de séjour nette ;
- Le type de cas ;
- l'âge à l'entrée ;
- La durée de séjour ;
- Le tarif ;
- La classe d'hospitalisation ;
- Les heures de soins intensifs ;
- Le groupe de classe ;
- Le sexe ;
- Le type de patient.

8 Conclusion

L'objectif principal de ce travail était de fournir aux comptables du RSV une méthode d'estimation des recettes, plus efficace que celle existant, pour les cas qui ne peuvent être facturés lors du bouclement comptable.

La solution proposée se nomme PrediRec ; c'est une application Web, déposée dans l'Intranet du Data Warehouse, permettant d'estimer les recettes des cas non codés en utilisant les outils de Data Mining. Les utilisateurs de PrediRec n'étant pas informaticiens, un des critères était également la simplicité de la solution sur le plan technique. Une application Web s'est avérée finalement le choix le plus approprié, surtout du fait qu'aucune installation « client » n'est nécessaire.

L'application offre la possibilité de mettre à jour les modèles d'analyse, de choisir les cas devant être provisionnés et d'exporter les résultats dans MS Excel. Elle permet en outre d'effectuer la simulation d'un cas fictif. Etant donné que PrediRec peut être utilisé pas des utilisateurs germanophones ou francophones, l'application est utilisable soit en langue allemande, soit en langue française.

La paramétrisation par l'utilisateur, l'export des résultats et la gestion de deux langues sont donc également des critères de départ qui ont été remplis dans l'application remise.

Au niveau méthodologique, deux critères ont servi à la détermination des modèles d'analyse, dans un but d'optimiser l'efficacité de la solution proposée. Tout d'abord, le coefficient de corrélation R2 entre la recette estimée et la recette réelle a été fixé au minimum à 0.80 ». Ensuite, la « Différence, en %, entre la somme totale estimée par rapport à la somme totale réelle » a été minimisée.

Les tests réalisés ont permis de satisfaire à ces deux exigences qualitatives.

En conclusion, tant les aspects méthodologiques (modèles, seuils qualitatifs) que techniques (simplicité, fonctionnalités, bilinguisme, paramétrisation, etc.) ont permis à ce projet de remplir les objectifs de départ. Nous pouvons donc affirmer que le « delivery » correspond aux attentes du client.

Point de vue personnel :
Je suis très satisfait de mon travail de diplôme dans lequel j'ai pu exploiter et renforcer mes connaissances des divers logiciels utilisés pour la conception de ce projet de Data Mining.

Au début, j'appréhendais ce travail car je n'avais jamais eu de formation, ni de connaissance sur les outils de Data Mining. Je me posais aussi la question : « Est-ce que c'est possible, à partir des informations disponibles, de créer un modèle d'analyse permettant d'estimer le montant d'une facture en partant de cas existants ? »

Les premières semaines de ce travail de diplôme furent consacrées à la recherche de documentation et à la compréhension du logiciel Visual Studio 2005 Business Intelligence, ainsi qu'à l'exécution des divers tutoriaux Microsoft. Ceux-ci étant extrêmement vagues, tant du point de vue des écrans utilisés que de la pertinence des résultats annoncés, il m'a fallu créer moi-même le tutorial sur la prise en main de Visual Studio 2005 Business Intelligence.

La phase de collecte me pris aussi plus de temps que prévu initialement car dans un projet de Data Mining, un cas analysé doit correspondre à une seule et unique ligne dans la table

d'apprentissage. Pour arriver à cette finalité, j'ai dû créer plusieurs vues agrégeant certaines variables comme les durées aux soins intensif ou les totaux facturés.

En fin de compte, le principal sujet de satisfaction a été de pouvoir, dans les délais impartis, rendre une solution correspondant aux attentes du client et qui, selon les retours d'informations obtenus, permettent de répondre à la demande en amenant une réelle plus-value.

Déploiement :
Durant le premier trimestre 2008, PrediRec sera accessible au Chef de la division « Finance et Controlling » du SZO afin que celui-ci puisse l'utiliser à des fins de test avec des cas non codés réels. A la suite de ces tests et d'une validation de la part de celui-ci, nous ouvrirons l'accès à PrediRec aux divers comptables et facturistes du RSV.

Il est aussi prévu d'enrichir l'application Web de diverses fonctions permettant d'affiner la sélection des cas non codés. Actuellement, cette étape doit être effectuée dans MS Excel.

Il est aussi prévu que l'utilisateur de PrediRec n'aie plus besoin d'entrer son nom d'utilisateur.

Elargissement du domaine d'étude, pistes de réflexion :
Sur la base de ce projet, nous pourrions utiliser les connaissances acquises et les méthodes d'analyses afin de, par exemple, prévoir l'évolution de certaines pathologies en Valais (p.ex. cancer, maladies cardio-vasculaires ou grippe saisonnière sur le court terme), ou encore permettre aux facturistes de repérer des lacunes dans la saisie de prestations. Un élargissement dans la prévision des plannings des blocs opératoires ou la prescription de médicaments sont également envisageables. Un lien avec des développements dans l'automatisation du codage est également tout à fait réalisable.

Ces quelques pistes de réflexion, centrées pour l'instant sur le domaine hospitalier, laissent présager de l'incroyable variété d'utilisations possibles des techniques et outils de Data Mining en entreprise.

9 Analyse du travail de diplôme

9.1 Connaissances acquises

Au cours de ce travail de diplôme, j'ai acquis les connaissances nécessaires à la conception et l'implémentation d'un projet de Data Mining au sein d'une entreprise.

Résumé des domaines abordés :
SQL Server 2005
Prise en main de SQL Server Management Studio
Conception de modèles d'analyse via le langage DMX
Installation d'un Serveur Analysis Services

Visual Studio 2005 Business Intelligence
Prise en main du logiciel
Connexion aux sources de données
Conception d'un modèle d'analyse
Exploration d'un modèle d'analyse
Utilisation d'un modèle d'analyse
Validation d'un modèle d'analyse

Data Mining
Etudes des modèles d'analyse
Choix des données d'analyse
Compréhension des algorithmes d'analyse

DMX
Prise en main du langage (création, modification, traitement et utilisation d'un modèle d'analyse)

ASP .NET
Prise en main du langage ASP.NET
Prise en main du langage AnalysisServices.AdomdClient

9.2 Problèmes rencontrés + solutions trouvées

Problème :
Le passage d'un modèle d'analyse créé avec Visual Studio 2005 Business Intelligence en solution interrogeable depuis un site Internet.

Solutions :
Reprise des différents tutoriaux Microsoft prévu pour l'interface Visual Studio 2005 Business Intelligence pour les refaire avec le langage DMX.

Problèmes :
Comment comparer les différents modèles d'analyses créés sans avoir accès à SSAS ou sans avoir à exécuter les prédictions lors des évaluations des modèles.

Solution :
La mise en place d'un classeur MS Excel.

En créant un classeur par modèle d'analyse testé, nous pouvions nous réunir sans avoir besoin d'un accès à notre serveur d'analyse.

Cette solution nous apporta un nouveau problème :
Comment choisir, parmi tous les résultats (une quarantaine), ceux qui doivent être analysé plus finement ?
La solution fut la mise en place d'un nouveau classeur MS Excel synthétisant tous les résultats stockés dans les autres classeurs MS Excel.

Ce nouveau fichier exécute les tâches suivantes :

- parcourir un répertoire
- ouvrir les fichiers MS Excel de ce répertoire
- copie des cellules précises du classeur de résultat dans le classeur de synthèse
- fermer les fichiers MS Excel de résultats

Problème :
Déploiement de l'application Internet PrediRec sur le serveur Internet du Data Warehouse. Lors des premiers tests, effectués après moult paramétrage de l'application Web, des composants nécessaires au bon fonctionnement du site Web étaient manquants et il était exclus d'utiliser une licence MS SQL Server 2005 sur ce serveur.

Solution :
Recherche sur Internet des composants nécessaires pour le fonctionnement d'un site Web utilisant la technologie « Microsoft.AnalysisServices.AdomdClient».

Après quelques recherches et l'installation d'un mauvais pack de composant, nous avons téléchargé et installé « SQLServer2005_ADOMD.msi » disponible à l'adresse suivante : http://www.microsoft.com/downloads/details.aspx?FamilyID=50B97994-8453-4998-8226-FA42EC403D17&displaylang=en.

Pour que l'application Web fonctionne correctement, il était nécessaire d'installer le « Feature Pack for Microsoft SQL Server 2005 » de février 2007 et non celui d'avril 2006.

9.3 Déclaration sur l'honneur

Je, soussigné Mathieu Giotta, déclare sur l'honneur :

- de ne pas distribuer ce dossier en dehors du cadre de mon travail de diplôme de l'HES-SO Valais ;
- n'avoir reçu aucune aide extérieure pour l'élaboration de ce travail de diplôme

Mathieu Giotta

9.4 Remerciements

Je remercie l'HEVs pour m'avoir laissé effectuer mon travail de diplôme dans l'entreprise pour laquelle je travaille.

Je remercie M. Gnaegi pour la proposition et le suivi en interne de ce travail de diplôme.

Je remercie M. Werlen pour m'avoir consacré du temps afin de m'expliquer comment étaient estimés les cas non codés lors des bouclements comptables les années précédentes.

Je remercie les codificatrices pour m'avoir accordé du temps afin de m'expliquer comment se passe la codification d'un dossier facturé par APDRG.

Je remercie les facturistes pour m'avoir accordé du temps afin de m'expliquer comment se passe la facturation au sein du CHCVs.

Je remercie M. Mueller pour m'avoir suivi durant la réalisation de ce travail de diplôme.

10 Glossaire

APDRG

Les DRG (Diagnosis Related Groups) sont des systèmes classant les séjours hospitaliers de soins aigus somatiques, sur la base des données récoltées de routine, dans un nombre défini de groupes homogènes du point de vue clinique et du point de vue de la consommation de ressources.

Il existe une multitude de systèmes cousins, dont les plus courants sont les Refined DRGs (RDRG), les All Patient DRGs (APDRG), les All Patient Refined DRGs (APRDRG) ou encore les International-Refined DRGs (IRDRG). La plupart des pays utilisent l'un ou l'autre de ces systèmes, ou ont développé leur propre système national (G-DRG en Allemagne, NordDRG en Scandinavie, ARDRG en Australie, etc.)

Les systèmes DRG ont été introduits aux États-Unis en 1983 pour le financement des soins. Utilisés dans la plupart des pays occidentaux, l'Allemagne et la France ont décidé récemment de leur introduction généralisée. Ces systèmes fournissent un outil performant pour la gestion de l'hôpital, en favorisant la rationalisation des investissements, une meilleure maîtrise des coûts et permettant la comparaison inter-établissements (benchmarking). L'intérêt supplémentaire de ce mode de remboursement est d'être basé sur les données médicales du patient et donc de tenir compte du coût du traitement, contrairement au forfait par jour traditionnel.

Introduits et testés en Suisse en 1998, les APDRG sont utilisés par un nombre croissant d'hôpitaux. Il est prévu qu'en 2006 près d'un hôpital sur deux facturera ses prestations sur cette base. Le projet de recherche et développement APDRG 1998-2004 avait pour but d'adapter cette technique aux besoins suisses et de réaliser sa mise en application. Les premières utilisations dans quelques cantons ces dernières années ont plus que confirmé les espoirs des initiateurs. Le nouveau club prend la relève de ce projet, pour en assurer la maintenance régulière et encourager la coopération entre ces utilisateurs.

Le succès du projet APDRG a incité les autorités et partenaires sanitaires de lancer un nouveau projet de recherche et de développement, SwissDRG 2004-2007. Ce projet a pour ambition de poursuivre la réflexion sur le financement des hôpitaux et d'identifier les solutions les plus judicieuses pour la Suisse dans le futur.

Carte de Kohonen

Les cartes de Kohonen ou SOM (Self-Organizing Maps) sont des réseaux de neurones à apprentissage non supervisé où l'on impose une topologie aux nœuds de sortie, en général sous forme de treillis rectangulaire ou hexagonal. Lorsqu'un nœud se déplace sous l'effet de l'apprentissage, il entraine automatiquement ses proches voisins, ce qui a pour conséquence que le treillis tente de reproduire la zone d'entrée. On peut voir par exemple des carrés qui essaient de

	se faire passer pour des cercles... Kohonen est donc le premier à avoir réussi la quadrature du cercle!
Centile	Un centile correspond au pourcentage d'individus de l'échantillon de normalisation qui ont obtenu un score inférieur à un score brut donné.
CHCVs	Acronyme de Centre Hospitalier du Centre du Valais. Centre hospitalier qui regroupe les hôpitaux de Sierre, Sion, Martigny et les cliniques du CVP et St-Claire.
CIM	Classification statistique Internationale des Maladies La CIM permet le codage des maladies, des traumatismes et de l'ensemble des motifs de recours aux services de santé. Elle est publiée par l'OMS et est utilisée à travers le monde pour enregistrer les causes de morbidité et de mortalité, à des fins diverses parmi lesquelles le financement et l'organisation des services de santé ont pris ces dernières années une part croissante.
Le codage	Transcription des diagnostics et des interventions décrits dans le dossier médical du patient, essentiellement dans la lettre de sortie et le rapport opératoire. Les diagnostics sont codés en Suisse selon la Classification statistique internationale des maladies et des problèmes de santé connexes, 10ème révision (CIM-10), publiée par l'OMS, alors que les interventions et les traitements sont codés selon la Classification suisse des interventions chirurgicale (CHOP), qui est une adaptation de la classification américaine ICD-9-CM, Volume 3. En Suisse, une nouvelle version de la CHOP est publiée chaque année par l'Office fédéral de la statistique. En 2007, c'est la version 9.0 qui a cours. L'obligation du codage faite aux hôpitaux et l'utilisation des classifications de référence figurent dans une annexe de l'Ordonnance fédérale du 30 juin 1993 concernant l'exécution des relevés statistiques fédéraux, en particulier la statistique médicale des hôpitaux. Cette ordonnance accompagne la Loi fédérale du 9 octobre 1992 sur la statistique fédérale, qui met sur pied et motive les différents relevés statistiques dans le domaine de la santé. La statistique médicale des hôpitaux, qui impose le codage dans tous les hôpitaux suisses, avait quatre buts principaux : - Surveillance épidémiologique (incidence et prévalence des maladies, état de santé de la population et mesures préventives ou thérapeutiques) - Saisie de prestations médicales homogènes et contrôle de la qualité - Bases pour la planification intra- et inter-cantonale - Mise à disposition de données pour la recherche et publications Avec l'introduction de la Loi fédérale sur l'assurance-maladie (LaMal) du 18.03.1994 et son Ordonnance sur le calcul des coûts et le classement des prestations par les hôpitaux et les établissements médico-sociaux dans l'assurance-maladie

	(OCP) de 2003, l'accent a surtout été mis sur le deuxième objectif. Ainsi, le codage médical a été la base du financement des hôpitaux par pathologie, selon le système APDRG (All Patient Diagnosis Related Groups), dès 2002 dans le canton de Vaud et dès 2004/2005 en Valais et dans d'autres cantons. La révision en cours de la LaMal prévoit d'ailleurs la généraliation d'un tel mode de financement dans tous les hôpitaux (SwissDRG) dès 2010/2011.
Comorbidité	En médecine, la comorbidité désigne : - la présence d'un ou de plusieurs troubles associés à un trouble ou une maladie primaire - l'effet provoqué par ces troubles ou maladies associés.
Cost-Weight (CW)	Chaque groupe de pathologie a son propre poids appelé cost weight (CW). Le cost weight indique le poids des frais de traitement moyens des patients d'un groupe DRG par rapport à celui de l'ensemble des patients en traitement stationnaire aigu en Suisse. Il est par exemple de 3,067 pour le groupe APDRG 191 (shunt intra-abdominal et interventions sur le pancréas et le foie, avec cc [comorbidités et/ou complications]) et de 1,432 pour le groupe APDRG 554 (interventions sur hernie, avec cc majeure). En d'autres termes, on suppose que le traitement d'un patient du groupe 191 coûte en moyenne 2,14 fois plus que celui d'un patient du groupe 554 et 3,067 fois plus que le traitement d'un patient moyen en traitement stationnaire aigu en Suisse.
Cost-Weight pondéré	Le Cost-Weight pondéré correspond au fait d'adapter le Cost-Weight de l'APDRG car, chaque APDRG possède une durée de séjour comprise entre une borne inférieure et une borne supérieure. Si la durée de séjour d'un cas n'entre pas dans cet intervalle, le CW est modifié soit à la hausse, soit à la baisse, afin d'ajuster les frais de traitements. (voir APDRG, Inliers, Outliers)
Data Warehouse	Structure informatique dans laquelle est centralisé un volume important de données consolidées à partir des différentes sources de renseignements d'une entreprise (notamment les bases de données internes). L'organisation des données est conçue pour que les personnes intéressées aient accès rapidement et sous forme synthétique à l'information stratégique dont elles ont besoin pour la prise de décision.
Data Mart	Sous ensemble d'un entrepôt de données, contenant des informations se rapportant à un secteur d'activité particulier de l'entreprise ou à un métier qui y est exercé (commercial, marketing, comptabilité, etc.).
Données d'apprentissage	Set de données utilisé dans un outils de Data Mining afin qu'il puisse les analyser.
DMX	Le langage DMX (Data Mining eXtensions) permet de créer et d'utiliser des modèles d'exploration de données dans Microsoft SQL Server 2005 Analysis Services (SSAS). Vous pouvez utiliser DMX pour créer la structure de nouveaux modèles d'exploration de données, pour l'apprentissage de ces modèles et pour les explorer, les gérer et y effectuer des prévisions. Le

	langage DMX est composé d'instructions de langage de définition de données (DDL, Data Definition Language), d'instructions de langage de manipulation de données (DML, Data Manipulation Language), ainsi que de fonctions et d'opérateurs.
EM	Expectation-maximisation L'algorithme espérance-maximisation), proposé par Dempster et al. est une classe d'algorithmes qui permettent de trouver le maximum de vraisemblance des paramètres de modèles probabilistes lorsque le modèle dépend de variables latentes non observables. On utilise souvent Espérance-maximisation pour la classification de données, en apprentissage machine, ou en vision artificielle. Espérance-maximisation alterne des étapes d'évaluation de l'espérance (E), où l'on calcule l'espérance de la vraisemblance en tenant compte des dernières variables observées, et une étape de maximisation (M), où l'on estime le maximum de vraisemblance des paramètres en maximisant la vraisemblance trouvée à l'étape E. On utilise ensuite les paramètres trouvés en M comme point de départ d'une nouvelle phase d'évaluation de l'espérance, et l'on itère ainsi.
ETL	Acronyme de Extraction Transformation Load Logiciel utilisé afin d'extraire, de transformer et de charger des données dans un Data Warehouse.
Gène de Mendel	Les gènes de Mendel correspondent a une étude entrepris par Gregor Mendel. Il fit des tests de croisement avec des petits poids et en conclura des lois de la génétiques.
Groupe Statistique Tarmed	Regroupement des prestations Tarmed pour les utiliser dans diverses statistiques. C'est un groupement interne au RSV.
Hospitalisation	Sont réputés traitements hospitaliers au sens de l'art. 49, al. 1, de la loi les séjours à l'hôpital d'une durée d'au moins 24 heures pour des examens, des traitements et des soins. Les séjours à l'hôpital de moins de 24 heures, au cours desquels un lit est occupé durant une nuit, ainsi que les séjours à l'hôpital en cas de transferts dans un autre hôpital ou en cas de décès, sont également réputés traitements hospitaliers.
Inliers	Dans une distribution, un cas est réputé Inlier s'il se situe à l'intérieur de deux bornes (seuils minimum et maximum). Pour les APDRG, c'est la distribution de la durée moyenne de séjour "théorique" (ALOS) par APDRG qui a servi à déterminer ces seuils. (voir APDRG, Outliers)
Outliers	Cas extrême, atypique dans une distribution. Un Outlier peut être "High-Outlier" ou "Low-Outlier", s'il se situe à l'extrémité supérieure, respectivement inférieure de la distribution. Le Cost-Weight de facturation sera pondéré pour tenir compte des jours au-dessous ou au-dessus des bornes de la distribution. Un Low-Outlier aura donc une facture inférieure au forfait de base et un High-Outlier un forfait supérieur, afin que la recette soit la plus proche possible des coûts effectifs. (voir APDRG, Inliers, Outliers)

Niveau Tarmed 1, 2, 3, 4, 5	Nomenclature officielle utilisée pour regrouper les prestations Tarmed selon le secteur médical dans lequel elles peuvent êtres saisies.
OLAP	OnLine Analytical Processing Technique d'analyse, élaborée en 1993 par E.F. Codd, un des créateurs des bases de données relationnelles, à la demande de la firme Arbor Software (devenue aujourd'hui Hyperion). L'objectif était de pouvoir sélectionner des données selon des critères multiples. Aujourd'hui, OLAP permet aux décideurs, en entreprise, d'avoir accès rapidement et de manière interactive à une information pertinente présentée sous des angles divers et multiples, selon leurs besoins particuliers. Très utilisés dans les secteurs de la banque, des télécommunications et de la grande distribution, les serveurs OLAP sont des outils opérationnels, qui permettent de valider une stratégie mise en oeuvre ou de vérifier des tendances. Ainsi, on pourra souhaiter examiner l'évolution des ventes d'un produit donné, dans une région géographique précise, au cours d'une saison donnée. Il suffira de préciser ces trois dimensions d'analyse au moteur OLAP. Les valeurs trouvées dans la base pourront être représentées sous la forme d'un cube. Si l'on avait souhaité examiner plus de trois critères ou dimensions, on parlerait alors d'hypercube.
PrediRec	Nom du projet, concaténation des termes Prediction et Recette
S.E.M.M.A.	Sample, Explore, Modify, Model, Assess L'acronyme SEMMA (échantillonne, explore, modifie, modélise, évalue) se rapporte au noyau du processus de conduite de l'exploitation de données. Commençant par un échantillon statistiquement représentatif de vos données, SEMMA rend facile l'application des techniques exploratoires statistiques, de visualisation, de choix et transformation des variables prédictives les plus significatives, de modélisation des variables pour prévoir des résultats, et de confirmation de l'exactitude d'un modèle.
SGBDR	Système de Gestion de Base de Données Relationnelle Le concept permet de stocker et d'organiser une grande quantité d'information. Les SGBD permettent de naviguer dans ces données et d'extraire (ou de mettre à jour) les informations voulues au moyen d'une requête.
SIMAV	Acronyme de Service d'Informatique Médicale et Administrative Valaisan. Il est charger de développer et gérer l'informatique des établissement sanitaires du canton du Valais.
SSAS	Acronyme de SQL Server Analysis Services Ensemble d'outils de Business Intelligence compris dans l'offre SQL Server 2005.
SZO	Acronyme de Spital Zentrum OberWallis. Correspond au centre hospitalier du haut-Valais qui regroupe les hôpitaux de Viège et de Brig.

11 Bibliographie

11.1 Ouvrage

[1] Lefébure R., Venturi G.
« Data Mining, Gestion de la relation client, Personnalisation de sites Web »,
Edition Eyrolles (2001)

[2] Burquier, B.,
« Business Intelligence avec SQL Server 2005, Mise en œuvre d'un projet décisionnel »,
Edition Dunod

11.2 Publication

Auteurs
AZE Jérome, LUCAS Noël, SEBAG Michèle
LRI, CNRS UMR 8623, Université Paris-Sud, 91405 Orsay, FRANCE

Titre de l'article
« Fouille de données visuelle et analyse de facteurs de risque médical »

Lien Internet
http://cat.inist.fr/?aModele=afficheN&cpsidt=15209108

Auteurs
Thomas M. Lehmann, Mark O. Güld, Thomas Deselaers, Daniel Keysers, Henning Schubert, Klaus Spitzer, Hermann Neyb, Berthold B. Wein

Titre de l'article
« Automatic categorization of medical images for content-based retrieval and data mining »

Lien Internet
http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6T5K-4FDJRV1-1&_user=10&_rdoc=1&_fmt=&_orig=search&_sort=d&view=c&_acct=C000050221&_version=1&_urlVersion=0&_userid=10&md5=7bd3e953ef3a25bbe1377e4dfc93d0cc

12 Liens Internet

Tous les liens Internet ci-dessous ont été testés en date du 12.12.2007.

APDRG
www.apdrgsuisse.ch

Cartes de Kohonen
http://membres.lycos.fr/jfbouzereau/kohonen.html

Centile
http://www.er.uqam.ca/nobel/r30574/PSY1282/C4P3.html

CIM
http://fr.wikipedia.org/wiki/Classification_internationale_des_maladies

Comorbidité
http://fr.wikipedia.org/wiki/Comorbidit%C3%A9

CW
http://www.zmt.ch/fr/stationaere_tarife/stationaere_tarife_apdrg/stationaere_tarife_apdrg_grundlageninformationen.htm

Data Warehouse
http://www.journaldunet.com/encyclopedie/definition/205/51/20/datawarehouse.shtml

Data Mart
http://www.journaldunet.com/encyclopedie/definition/341/51/20/datamart.shtml

Data Mining
http://www.journaldunet.com/encyclopedie/definition/204/51/20/datamining.shtml

Définition Data Mining
http://www.journaldunet.com/encyclopedie/definition/204/51/20/datamining.shtml

DMX
http://technet.microsoft.com/fr-fr/library/ms132058.aspx

EM
http://fr.wikipedia.org/wiki/Algorithme_esp%C3%A9rance-maximisation

Hospitalisation
http://www.admin.ch/ch/f/rs/832_104/a3.html

Méthode de discrétisation
http://technet.microsoft.com/fr-fr/library/ms174512.aspx

OLAP
http://www.journaldunet.com/encyclopedie/definition/575/51/20/online_analytical_processing_i_olap_i.shtml

R2
http://www.modalisa.com/Lexique/R2coefdetermi.html
http://www.modalisa.com/Lexique/R2coefideter.html

S.E.M.M.A.
http://fr.wikipedia.org/wiki/SEMMA

SGBDR
http://fr.wikipedia.org/wiki/SGBDR

SIMAV
http://www.rsv-gnw.ch/index.php?option=com_content&task=view&id=102&Itemid=138&lang=fr

SZO + CHCVs
http://www.rsv-gnw.ch/

TARMED
http://www.tarmedsuisse.ch/

Les principaux logiciels de Data Mining :
Microsoft® SQL Server Analysis Services
http://technet.microsoft.com/fr-fr/library/ms175609.aspx

SAS® Enterprise Miner
http://www.sas.com/offices/europe/france/software/documents/brochure_em.pdf

SPSS® Clementine
http://www.spss.com/fr/clementine/

Spécifications SSAS
Les types de données
http://technet.microsoft.com/fr-fr/library/ms174796.aspx

Les types de contenu
http://technet.microsoft.com/fr-fr/library/ms174572.aspx

Algorithme Microsoft Association
http://technet.microsoft.com/fr-fr/library/ms174916.aspx

Clusters Microsoft
http://technet.microsoft.com/fr-fr/library/ms174879.aspx

Microsoft Decision Trees
http://technet.microsoft.com/fr-fr/library/ms175312.aspx

Microsoft Naive Bayes
http://technet.microsoft.com/fr-fr/library/ms174806.aspx

Microsoft Neural Network
http://technet.microsoft.com/fr-fr/library/ms174941.aspx

Microsoft Sequence Clustering
http://technet.microsoft.com/fr-fr/library/ms175462.aspx

Microsoft Time Series
http://technet.microsoft.com/fr-fr/library/ms174923.aspx

Visionneuse d'arborescences Microsoft
http://technet.microsoft.com/fr-fr/library/ms174503.aspx

Projet SIMAV

Tutorial SSAS &
Visual Studio 2005 Business Intelligence

Facturation par APDRG :
prédiction des recettes des cas non codés

Institut Central des Hôpitaux Valaisans
Av. Grand Champsec 86
Case postale 736
1951 Sion
Suisse

Auteur	: Mathieu Giotta	**Date de création**	: 08.11.2007
Fichier	: PrediRec - 003 Tutorial BI	**No de version**	: V. finalisée
Etat	:	**Dernière révision**	:
Distribution	:	**Date de distribution**	:
Publication	:		

Table des matières

1 Installation Microsoft SQL Server 2005 Analysis Services

Voici la procédure à suivre afin d'installer SQL Server 2005 Analysis Services, ci après SSAS.

Une fois le CD/DVD ROM de SQL Serveur 2005 inséré dans le lecteur optique, l'écran suivant devrait apparaître. (Figure 1-1)
Si ce n'est pas le cas, il suffit de double cliquer sur l'icône du lecteur optique.

Figure 1-1 : Menu « Démarrer » du CD-ROM

A l'écran de la Figure 1-1, il faut choisir, à partir du menu « Installer », le sous-menu « Compostant Serveur, documentation en ligne et exemples » en cliquant simplement dessus.

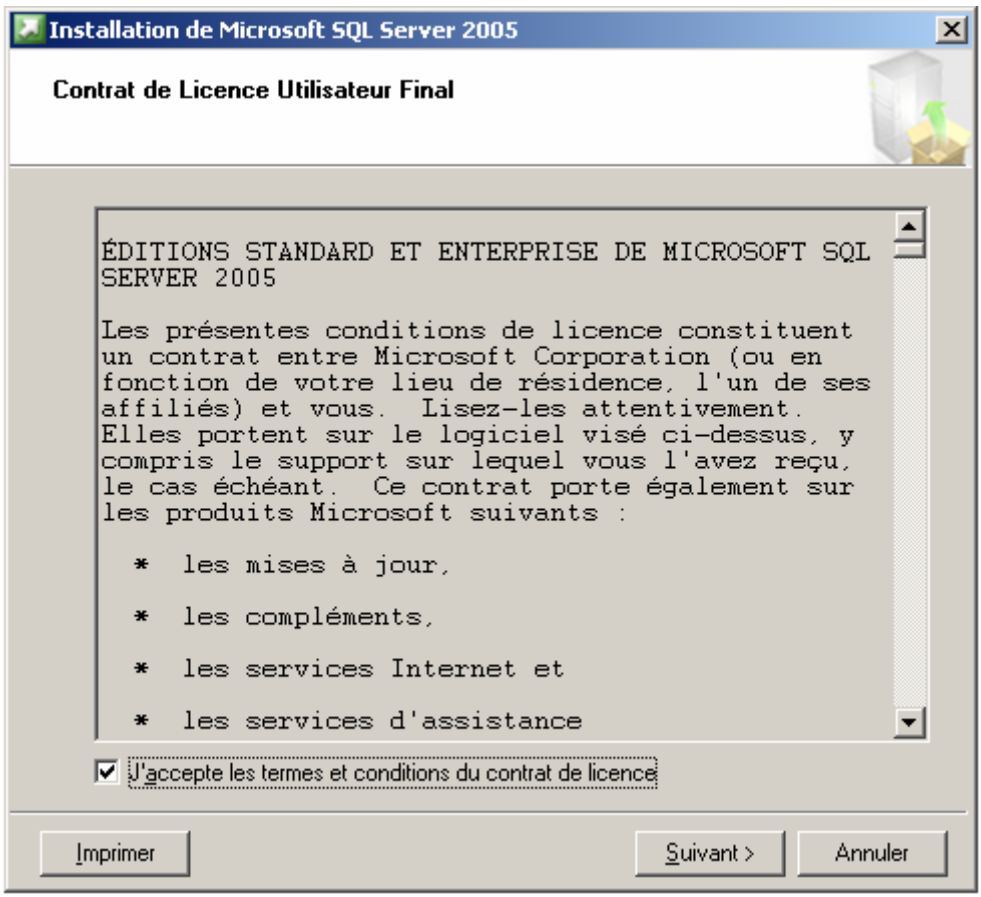


Figure 1-2 : Contrat de Licence Utilisateur Final

L'acceptation du contrat de Licence Utilisateur Final (Figure 1-2) doit être validée avant de pouvoir commencer l'installation. Pour le faire, il suffit de cocher la case « J'accepte les termes et conditions du contrat de licences » et de cliquer sur « Suivant ».

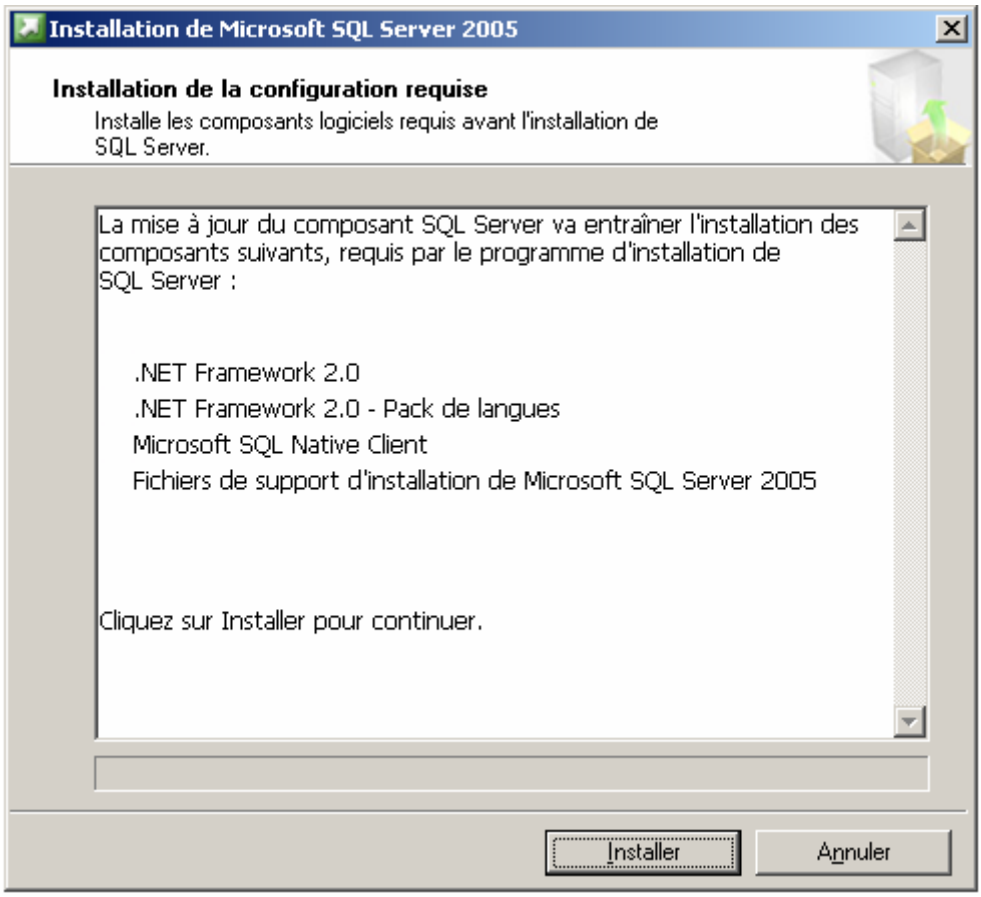


Figure 1-3 :Installation de la configuration requise

Avant de commencer l'installation de SSAS, l'installateur doit installer le .NET Framework 2.0 ainsi le pilote Microsoft SQL Server 2005 (Figure 1-3). Cliquer sur « Installer »
Après cette phase d'installation, 3 écrans d'informations se succèderons. Pour Chacun d'eux, il faut cliquer sur « Suivant ».

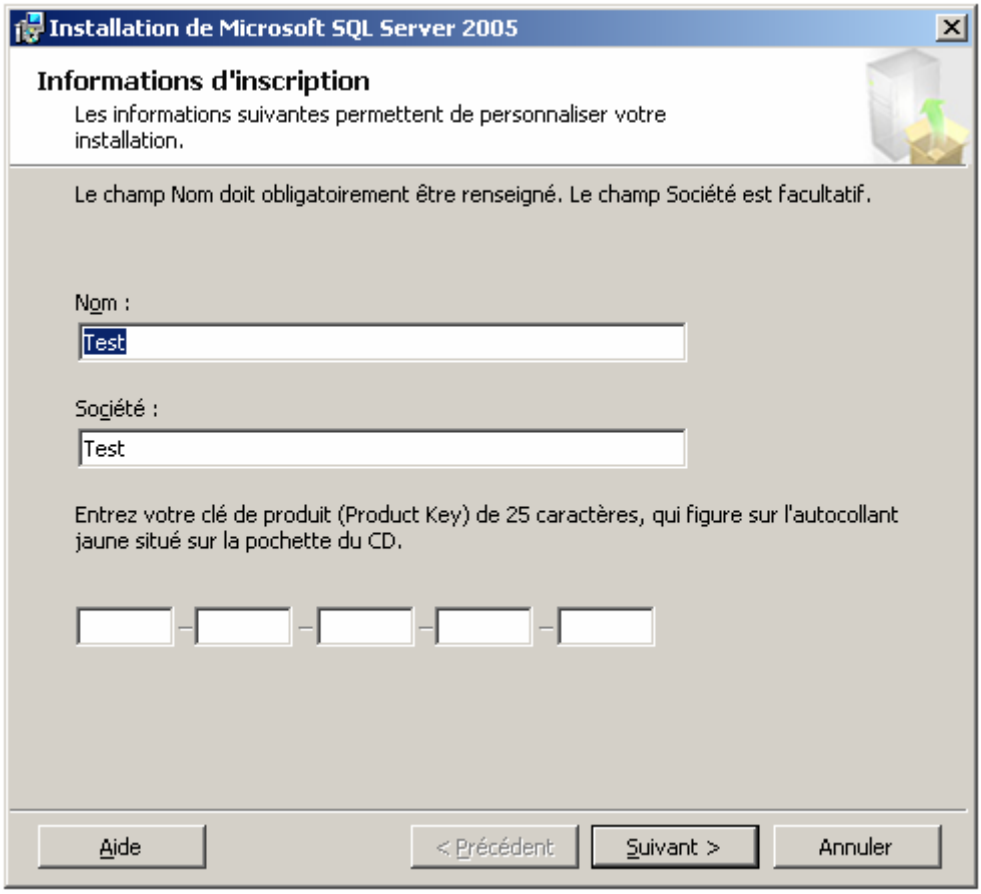


Figure 1-4 : Information d'inscription

Il est primordial, durant l'installation, de donner des informations nécessaire à l'inscription du produit, tel que le code d'activation (Figure 1-4).
Une fois les informations requises saisies, cliquer sur « Suivant ».

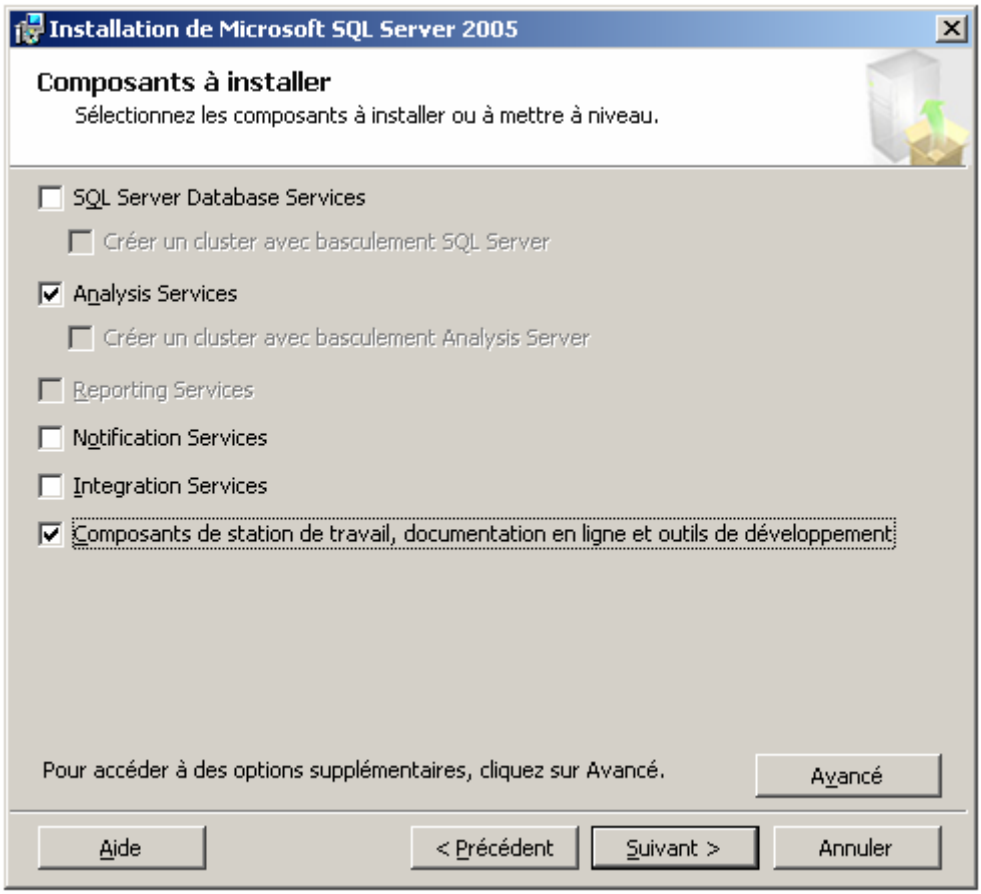


Figure 1-5 : Composant à installer

Etant donné que nous nous intéressons uniquement à l'installation de SSAS, nous cocherons uniquement les composants suivants à installer : « Analysis Services » et les « Composants de station de travail, documentation en ligne et outils de développement » (Figure 1-5). Une fois les cases cochées, cliquer sur « Suivant ».

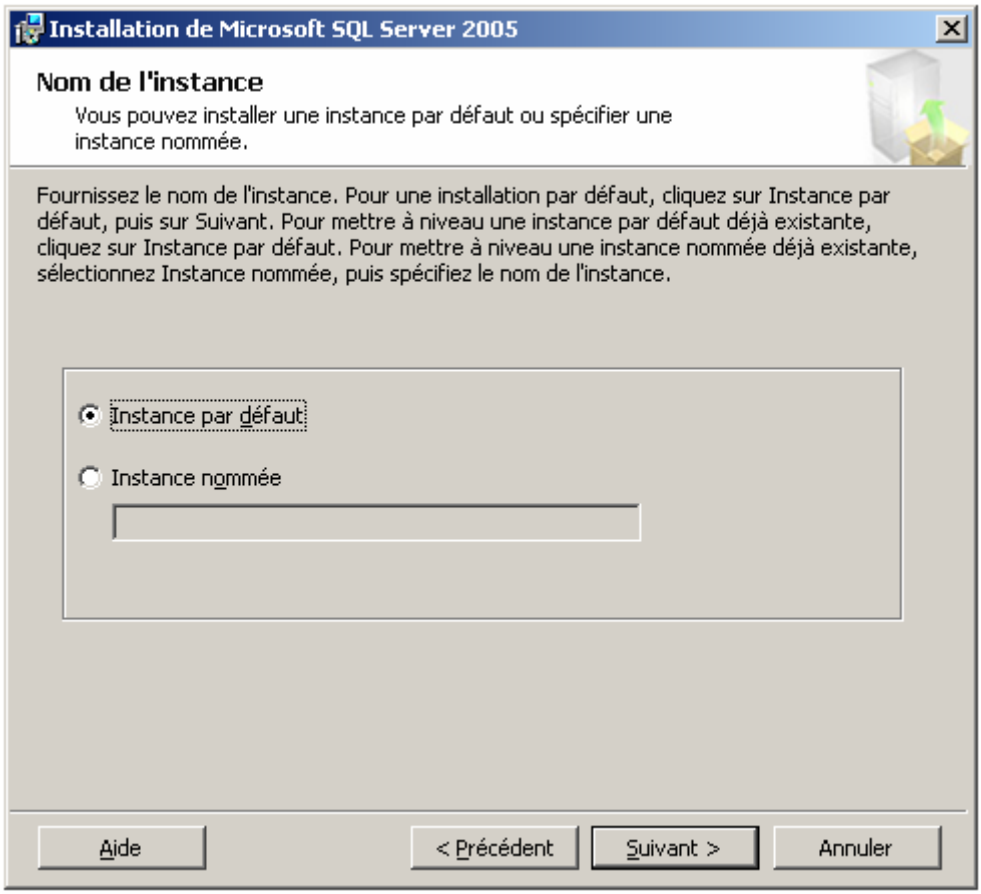


Figure 1-6 : Nom de l'instance

Pour cette installation, nous laissons le nom de l'instance par défaut comme sur la Figure 1-6. Celui-ci représente l'identité de ce serveur à travers le réseau de l'entreprise. Si besoin est, il est possible de nommer l'instance différemment.
Une fois l'option choisie, cliquer sur « Suivant ».

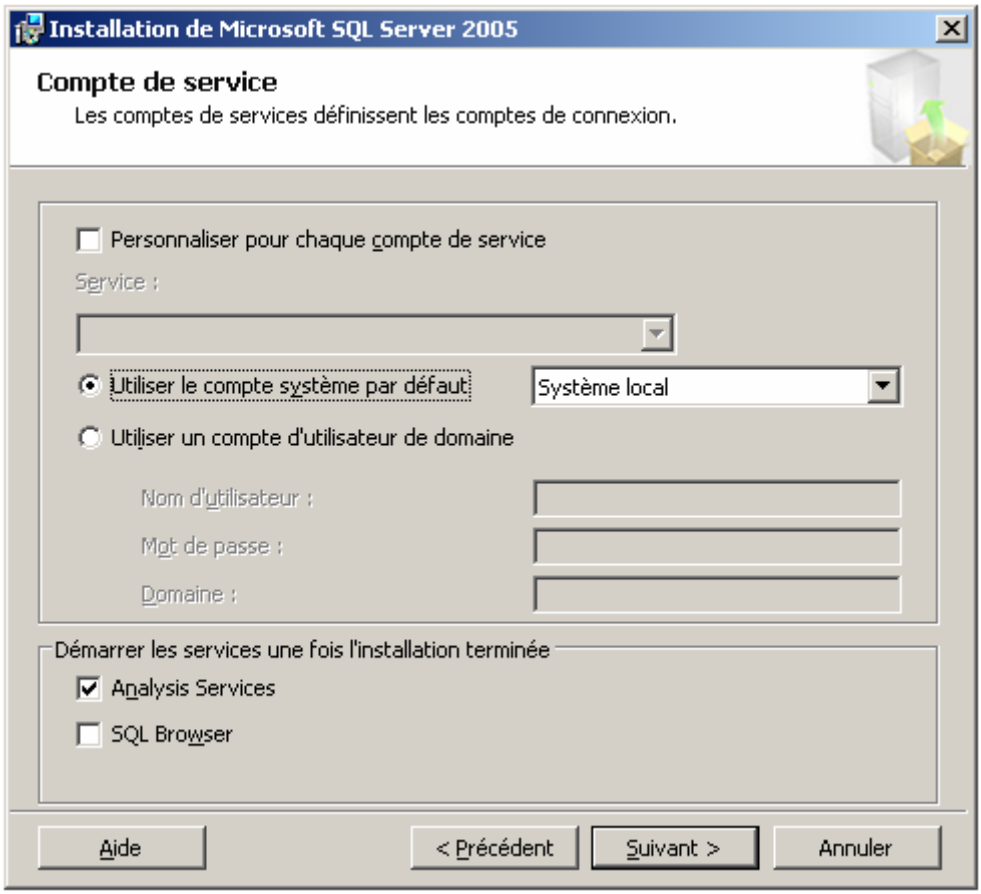


Figure 1-7 : Compte de service

Pour cette installation, nous configurons le « Compte de service » en sélectionnant « Utiliser le compte système par défaut » et nous laissons les autres options tel que proposées. (Figure 1-7). Cliquer sur « Suivant ».

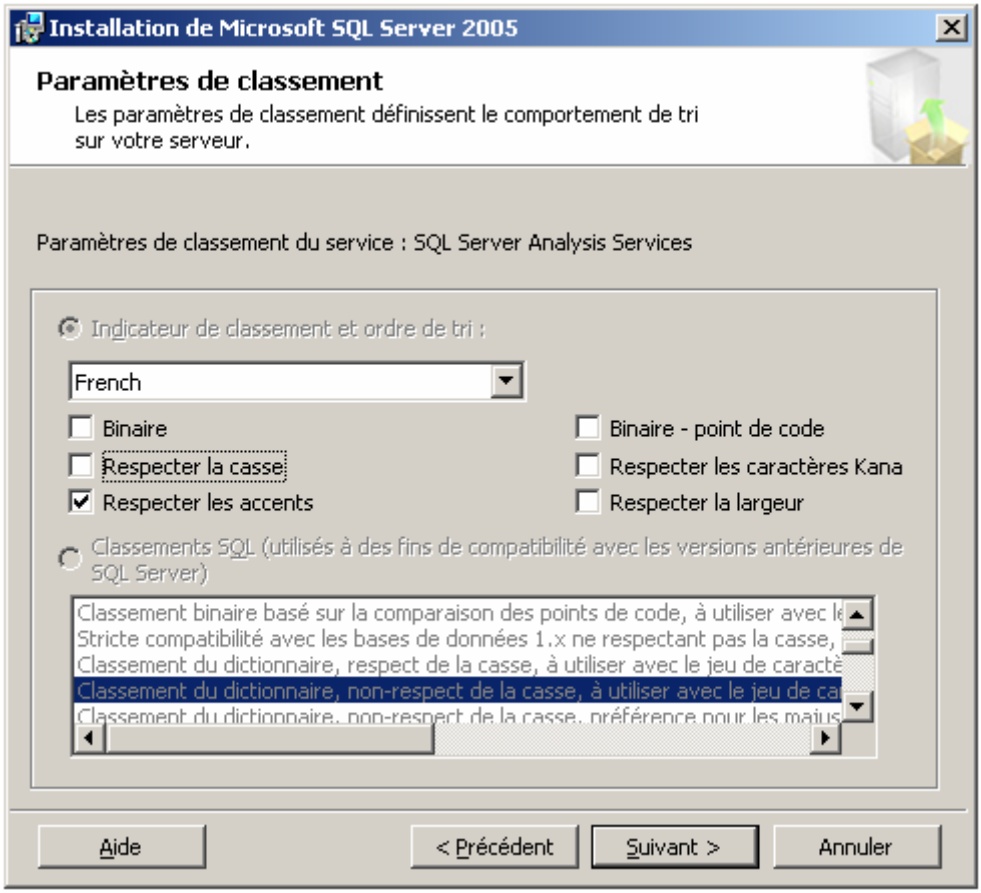


Figure 1-8 : Paramètre de classement

A l'écran « Paramètres de classement » (Figure 1-8), nous laissons les choix prédéfinis et nous cliquons sur « Suivant ».

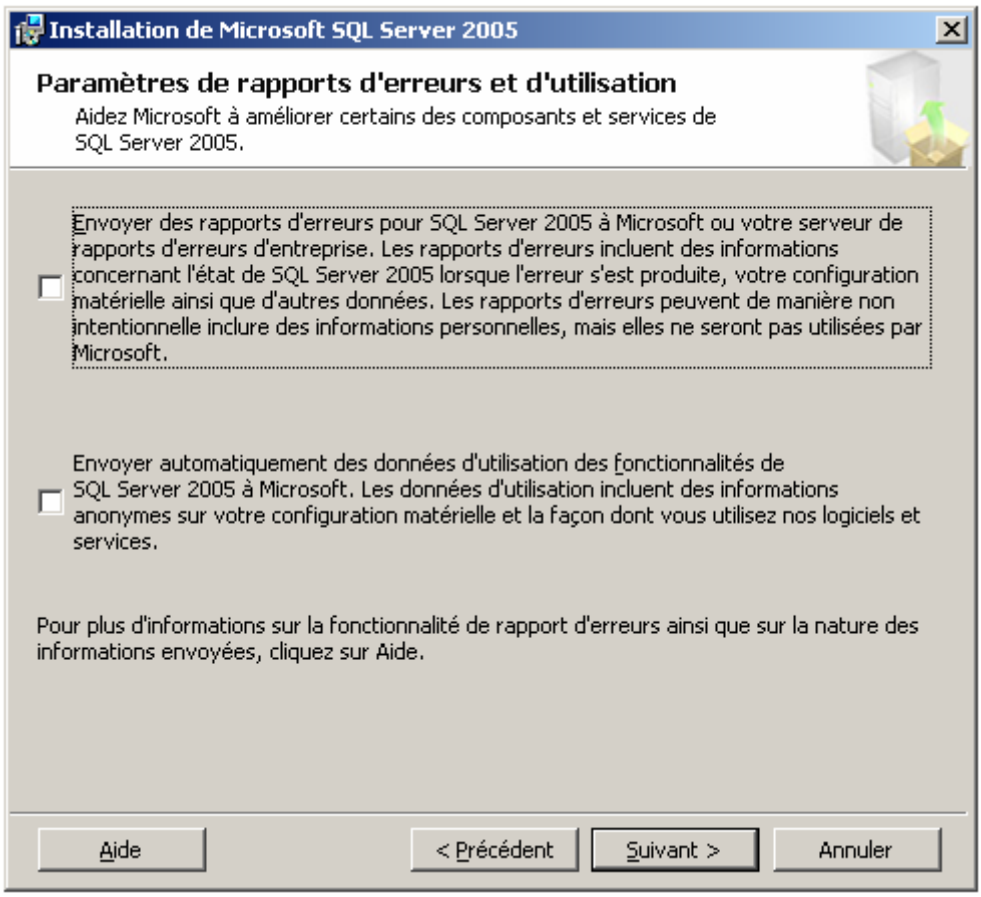


Figure 1-9 : Paramètres de rapports d'erreurs et d'utilisation

A l'écran « Paramètre de rapports d'erreurs et d'utilisation » (Figure 1-9), nous décochons toutes les options sélectionnées et nous cliquons sur « Suivant ».

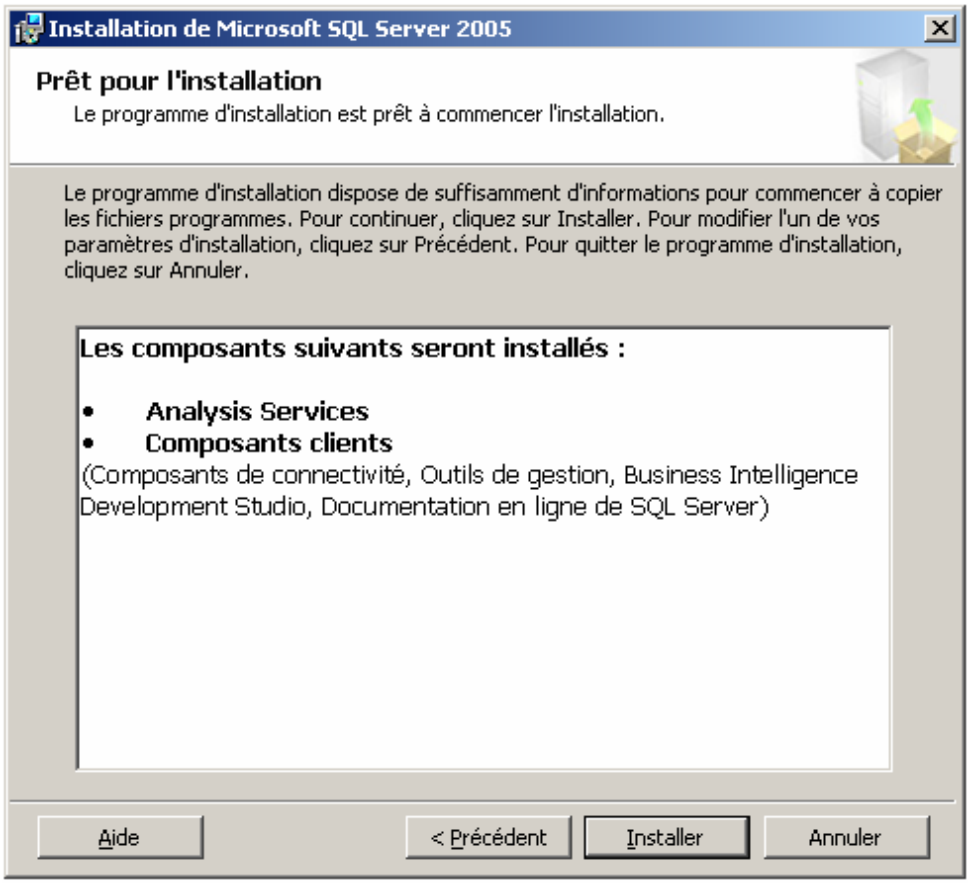


Figure 1-10 : Prêt pour l'installation

Une fois ces démarches finies, l'installation réelle peut commencer. Il ne reste plus qu'à cliquer sur « Installer » de l'écran « Prêt pour l'installation » (Figure 1-10).

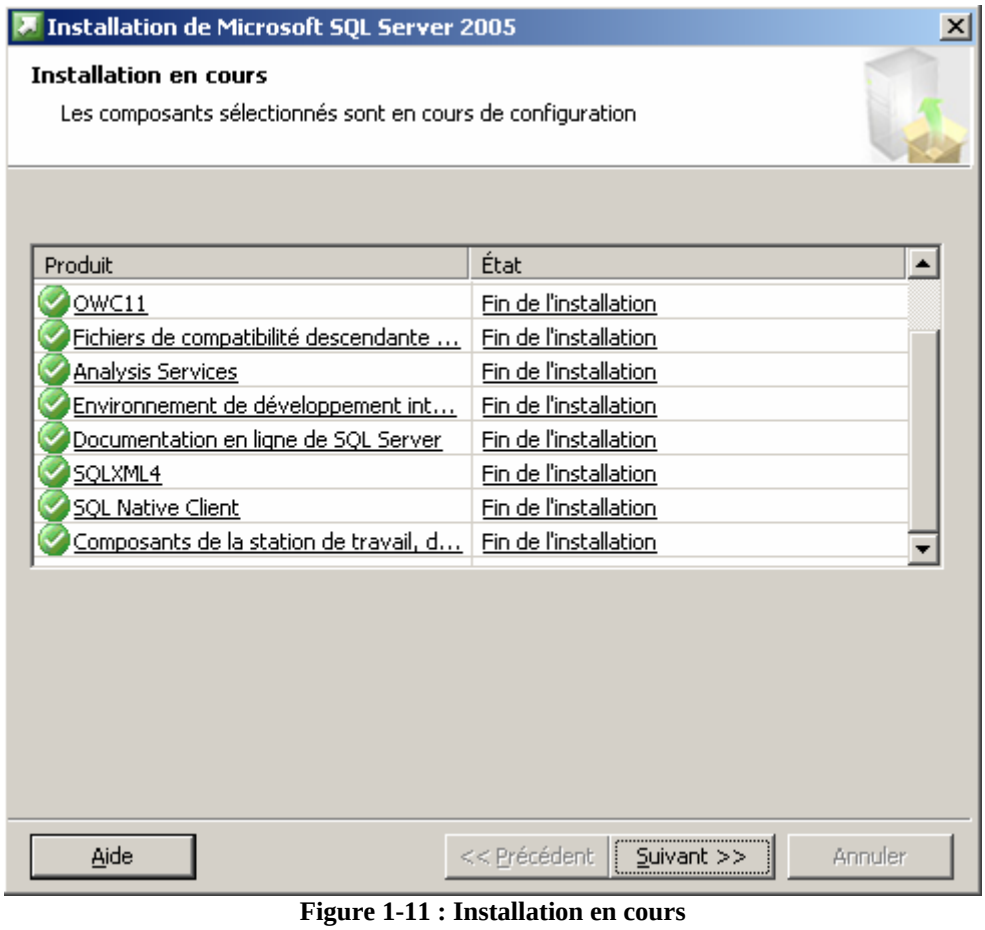


Figure 1-11 : Installation en cours

A la fin de l'installation, cliquer sur « Suivant »de l'écran « Installation en cours » (Figure 1-11). Nb. Ce bouton est grisé tant que l'installation n'est pas terminée.

Figure 1-12 : Fin de l'installation

Pour terminer l'installation, cliquer sur « Terminer » de la Figure 1-12.
A la suite de cette installation, je conseille de redémarrer l'ordinateur, même si le programme d'installation ne le propose pas..

Suite au redémarrage de l'ordinateur, vous aurez 2 nouveaux répertoires dans le menu démarrer (Figure 1-13) :

- Microsoft SQL Server 2005 ;
- Microsoft Visual Studio 2005.

Nb. Nous supposons que le serveur était au préalablement installer à neuf.

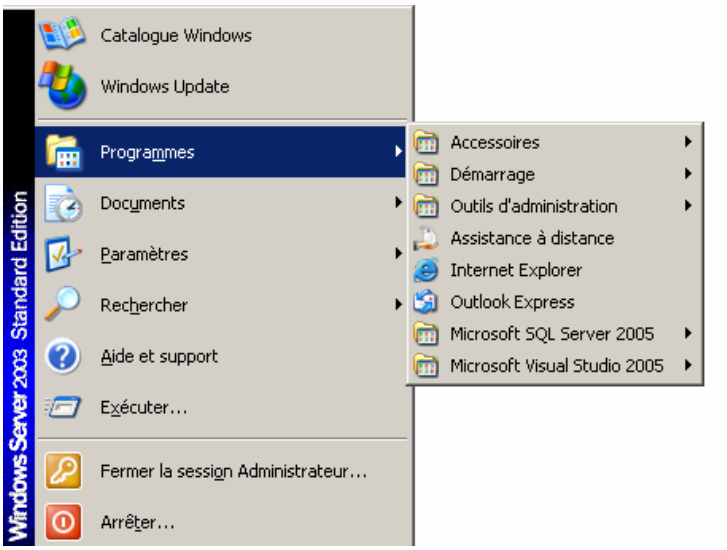


Figure 1-13 : Nouveau Menu Démarrer

Arrivés à cette étape, nous avons terminé d'installer nos services SSAS.

Le monde de la Business Intelligence (BI) s’ouvre à nous.

2 SSAS, premier contact

Il existe deux moyens différents pour accéder à SSAS :

- soit via SQL Server Management Studio (Figure 2-1)

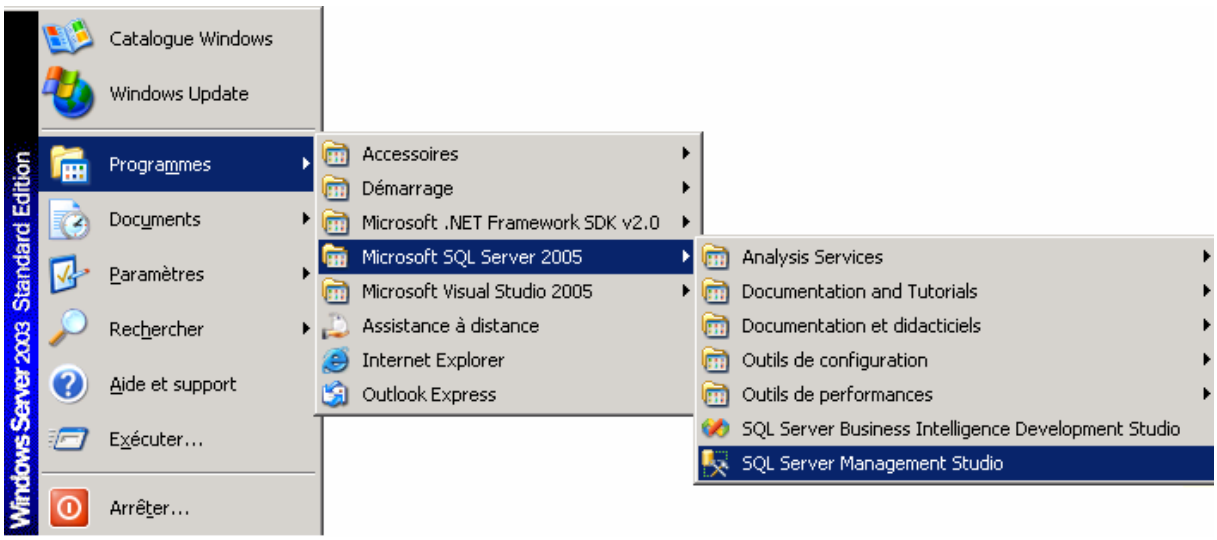


Figure 2-1 : Menu Démarrer, SQL Server Management Studio

- soit via Microsoft Visual Studio 2005 (Figure 2-2)

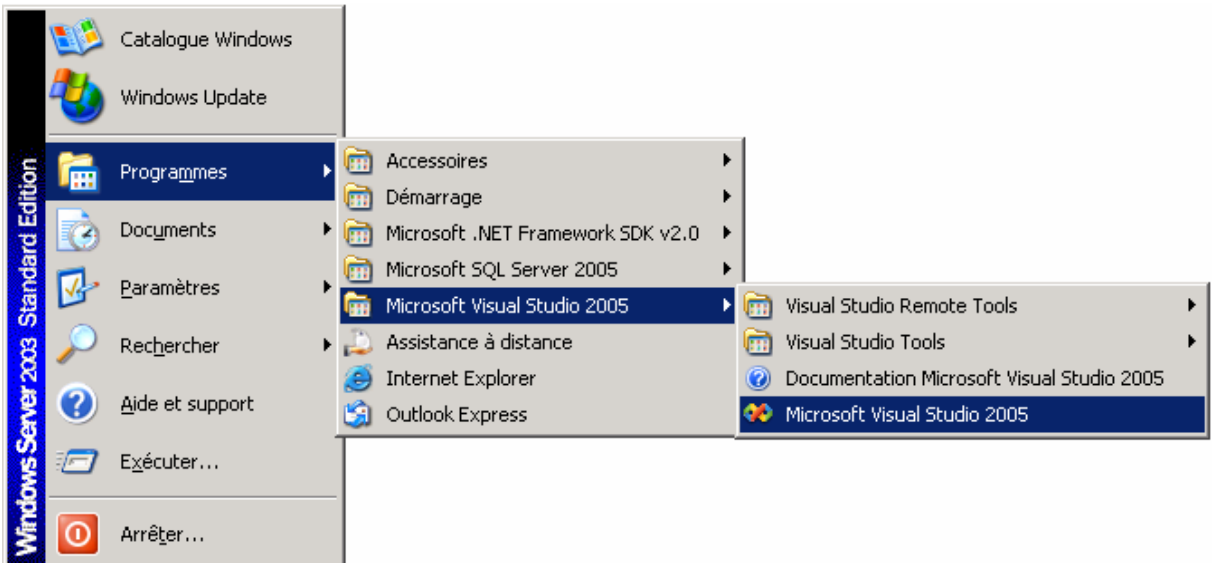


Figure 2-2 : Menu Démarrer, Microsoft Visual Studio 2005

Bien que pour la suite de ce tutorial nous utiliserons Microsoft Visual Studio 2005, je ferai une brève présentation de SQL Server Management Studio SSAS à la fin dudit tutorial.

Il est nécessaire de noter que si nous utilisons l'un ou l'autre programme, le fonctionnement reste le même.

Seuls l'interface et le maniement sont différents, mais tous deux utilisent le même moteur de Data Mining.

3 Microsoft Visual Studio 2005 – BI – tutorial

Pour ce tutorial l'interface utilisée est, comme annoncé, Microsoft Visual Studio 2005 – BI, ci-après VS2005 – BI.

Pré-requis :
Avant de commencer ce tutorial, il est nécessaire que la base de données AdventureWorks-DW soit installée sur votre serveur de base de données et que certaines corrections sur celle-ci doivent être effectuées.
L'annexe 6.1 indique la marche à suivre pour installer les bases de données de test de Microsoft SQL Serveur 2005.
L'annexe 6.2 contient divers script SQL (Script 6-1 à Script 6-4) qui corrigent certaines données de la base de test « AdventureWorks DW ».

Scénario :
Nous reprenons le tutorial proposé par Microsoft en expliquant beaucoup plus clairement certains points important et aussi en détaillant les écrans auxquels nous seront confrontés.

Le scénario proposé est de créer un publipostage ciblé. C'est-à-dire, qu'afin de limiter les coûts liés à l'envoi de publicité, notre entreprise de vente de vélos désire envoyer des flayers directement aux clients qui pourraient potentiellement acheter un vélo.

Pour effectuer ce publipostage, nous disposons d'une base de données clients contenant leurs informations personnelles ainsi que leurs habitudes d'achat.

De plus, nous partons sur l'hypothèse que la base de données clients ne comporte aucune erreur (manque d'information) sur les personnes y figurant.
Dans le cas contraire, il faudrait effectuer diverses tâches SQL Serveur Integration Services (SSIS) afin de « nettoyer » les données sources.

Tutorial :
Après avoir exécuté VS 2005 via le menu Démarrer, Programmes, Microsoft Visual Studio 2005, Microsoft Visual Studio 2005, nous allons démarrer un nouveau projet via le menu Fichier\Nouveau\Projet… (Figure 3-1)

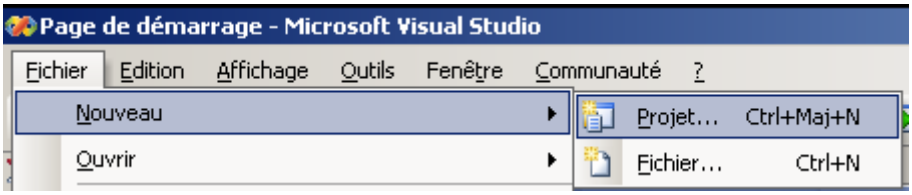


Figure 3-1 : Fichier, Nouveau, Projet…

A partir de l'écran « Nouveau projet » nous choisissons « Projet Analysis Services » que nous nommons « Tutorial Data Mining » et que nous validons en cliquant sur « OK » (Figure 3-2)

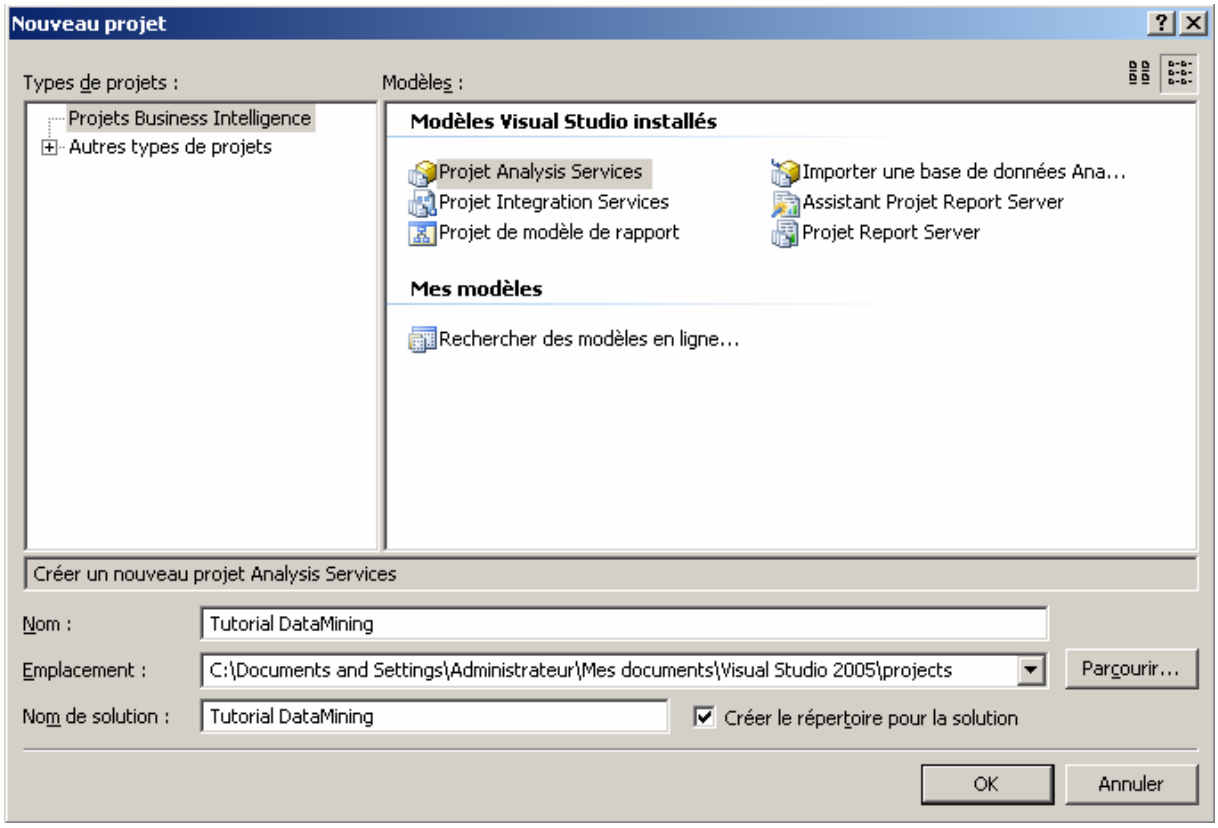


Figure 3-2 : Fenêtre « Nouveau projet » - Choix d'un « Projet Analysis Services »

A la suite de la création par VS 2005 de notre projet, nous pouvons apercevoir un explorateur de solution sur la droite de l'écran qui diffère des projets créés habituellement avec VS 2005 (Figure 3-3).

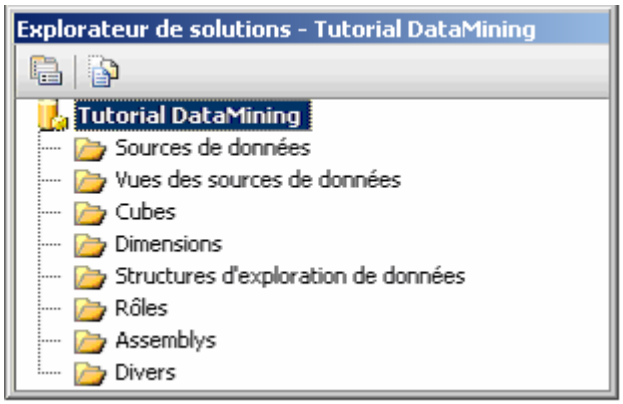


Figure 3-3 : Fenêtre : « Explorateur de solutions » - vide

Etant donné que l'outil Data Warehouse de Microsoft est également VS 2005 BI, cet explorateur contient plusieurs répertoires que nous n'utilisons pas dans ce tutorial.
Les dossiers que nous employons sont :

- Sources de données ;
- Vues des sources de données ;
- Structure d'exploration de données.

3.1 Créer une Sources de données

Une source de données représente une connexion sur un serveur de base de données

La première étape d'un projet SSAS Data Mining consiste à créer une source de données qui nous permet de nous connecter à nos données.

Pour ajouter une source de données, il suffit de faire un clic droit sur le dossier « Sources de données » et de choisir dans le menu « Nouvelle source de données » (Figure 3-4).

Figure 3-4 : Fenêtre « Explorateur de solutions » - Ajout Nouvelle source de données…

L'assistant « Source de données » s'affiche. Cliquer sur Suivant pour configurer une source de données (Figure 3-5).

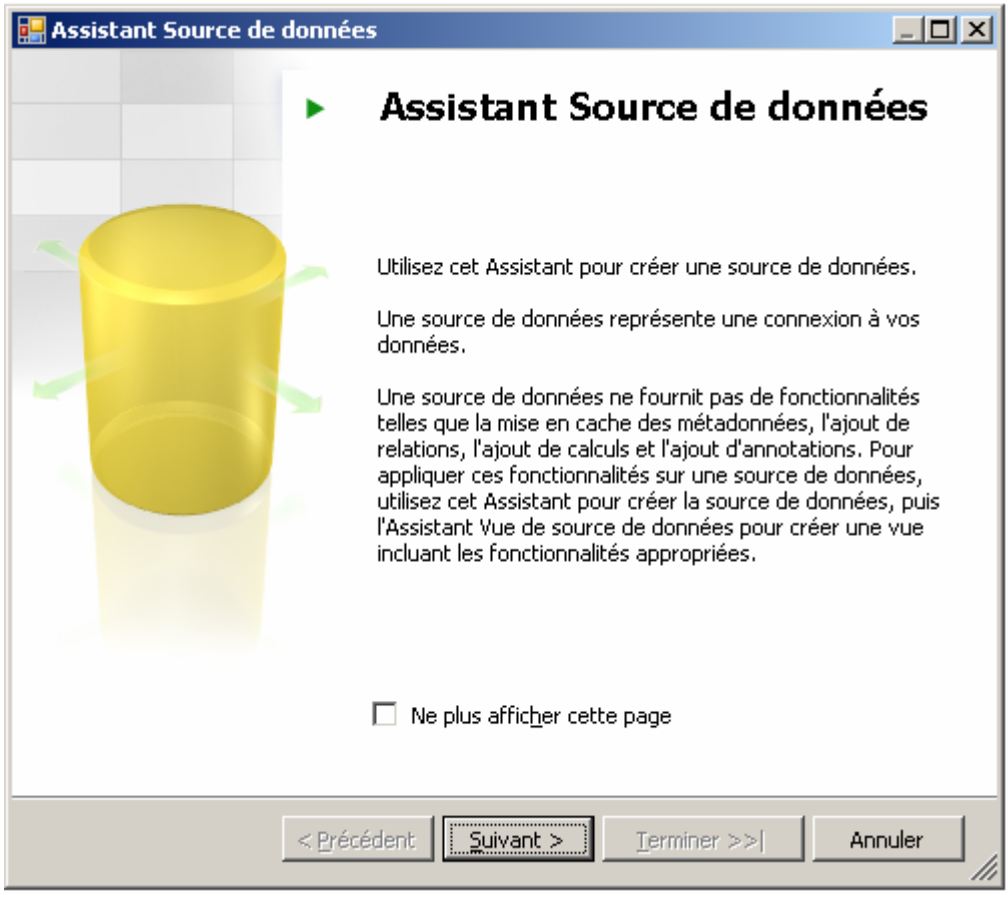


Figure 3-5 : Assistant Source de données

Sur l'écran « Sélectionner la méthode de définition de la connexion » (Figure 3-6), choisir l'option « Créer une source de données basée sur une connexion existante ou nouvelle » et ensuite cliquer sur « Nouveau… ».

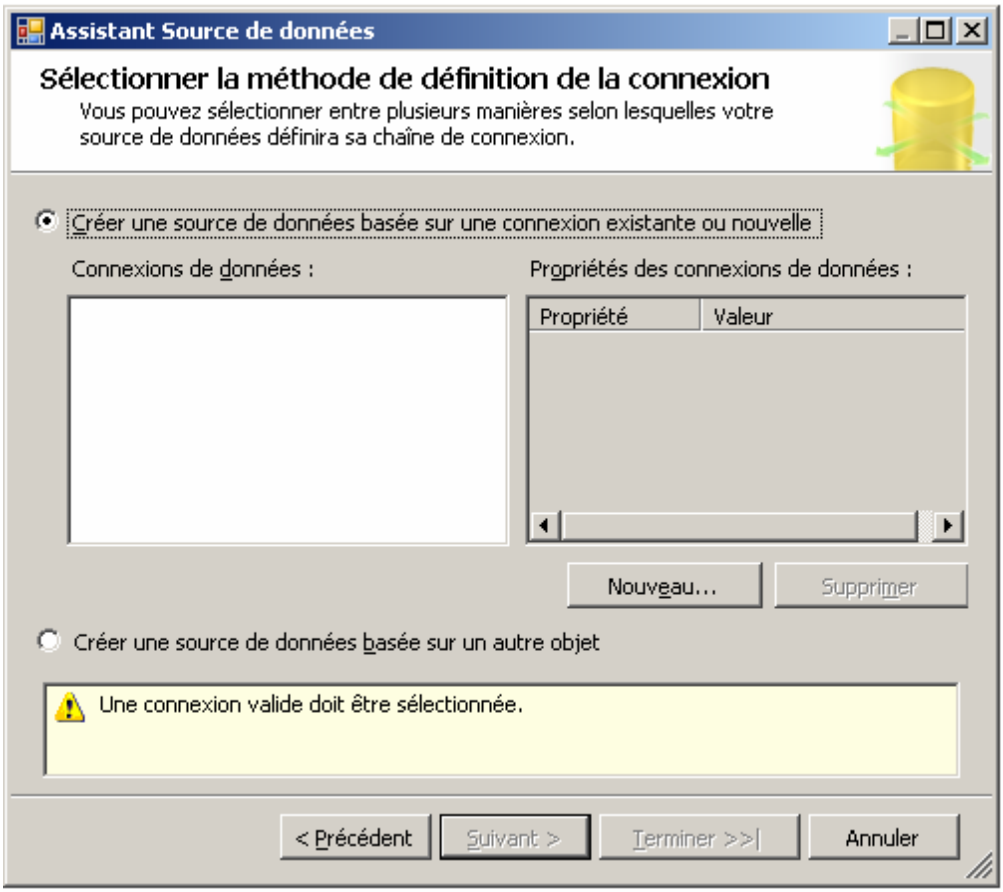


Figure 3-6 : Assistant Source de données - Sélectionner la méthode de définition de la connexion

Sur l'écran « Gestionnaire de connexion » (Figure 3-7), nous indiquons le Nom du serveur de base de données ainsi que la base de données voulues (ici AdventureWorksDW).
Des paramètres d'authentification peuvent être demandés selon la configuration du serveur de base de données.

Une fois les informations nécessaires saisies, valider en cliquant sur OK.

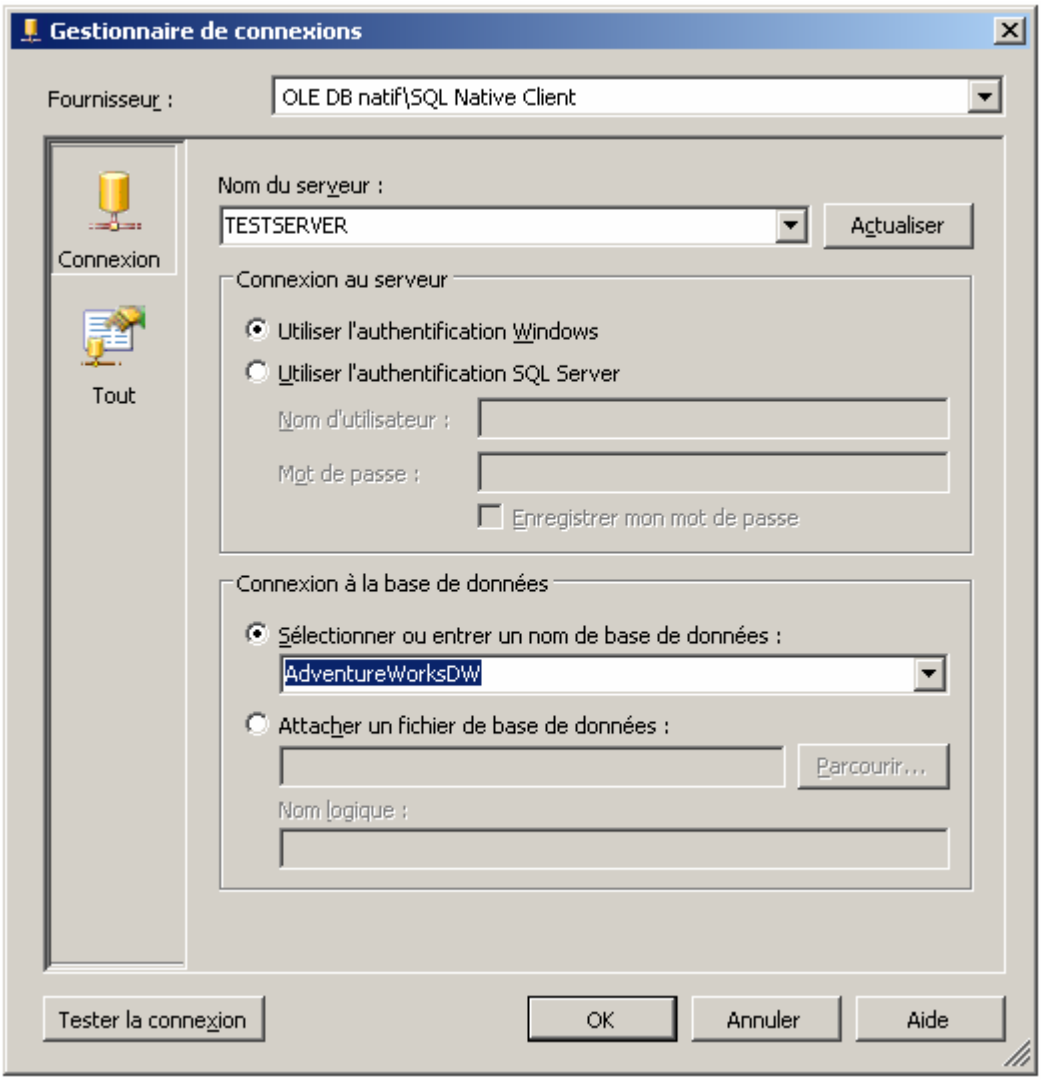


Figure 3-7 : Assistant Source de données – Gestionnaire de connexions

Nous sommes de retour à l'écran « Sélectionner la méthode de définition de la connexion » (Figure 3-8).

Dans certain cas, il peut être nécessaire de créer d'autres sources de données, mais pour ce tutorial, une seule source de données est utile. Cliquer sur « Suivant ».

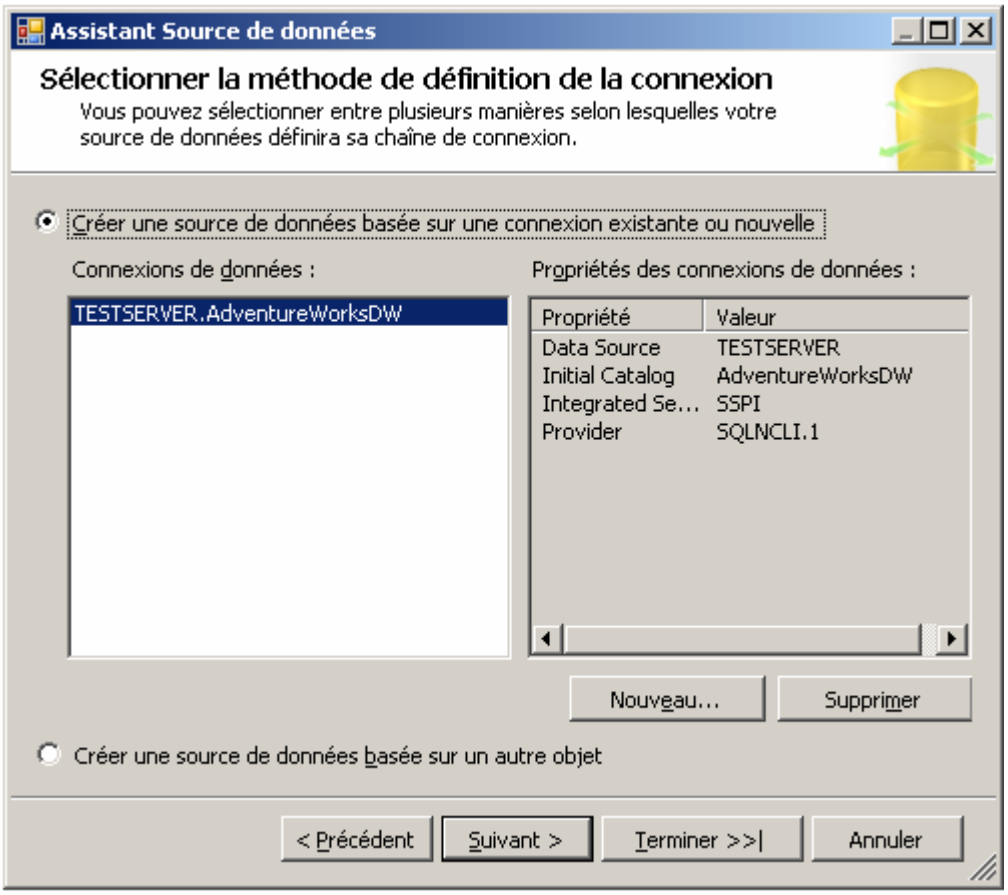


Figure 3-8 : Assistant Source de données

A l'écran « Assistant d'emprunt d'identité » (Figure 3-9) nous choisissons l'option « Utiliser le compte de service ». Cliquer sur « Suivant ».

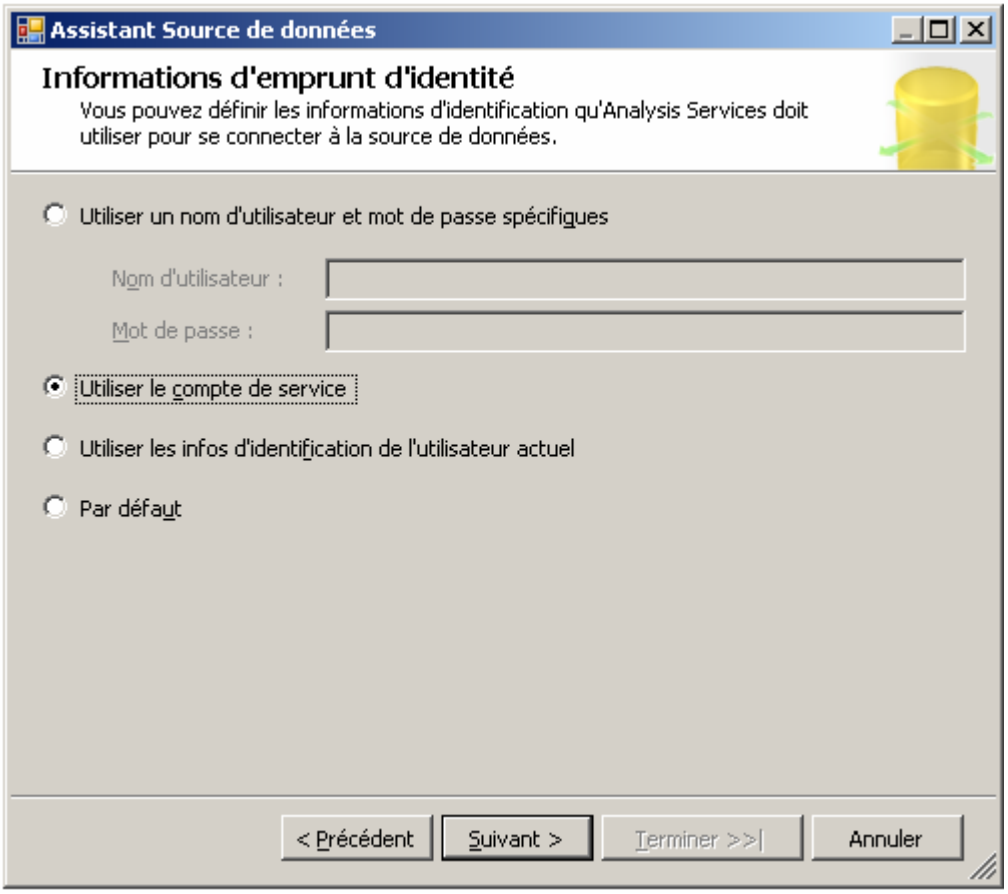


Figure 3-9 : Assistant Source de données - Information d'emprunt d'identité

Pour terminer, nous saisissons un nom pour identifier la source de données à l'écran « Fin de l'assistant » (Figure 3-10).
Pour achever l'installation, cliquer sur « Terminer ».

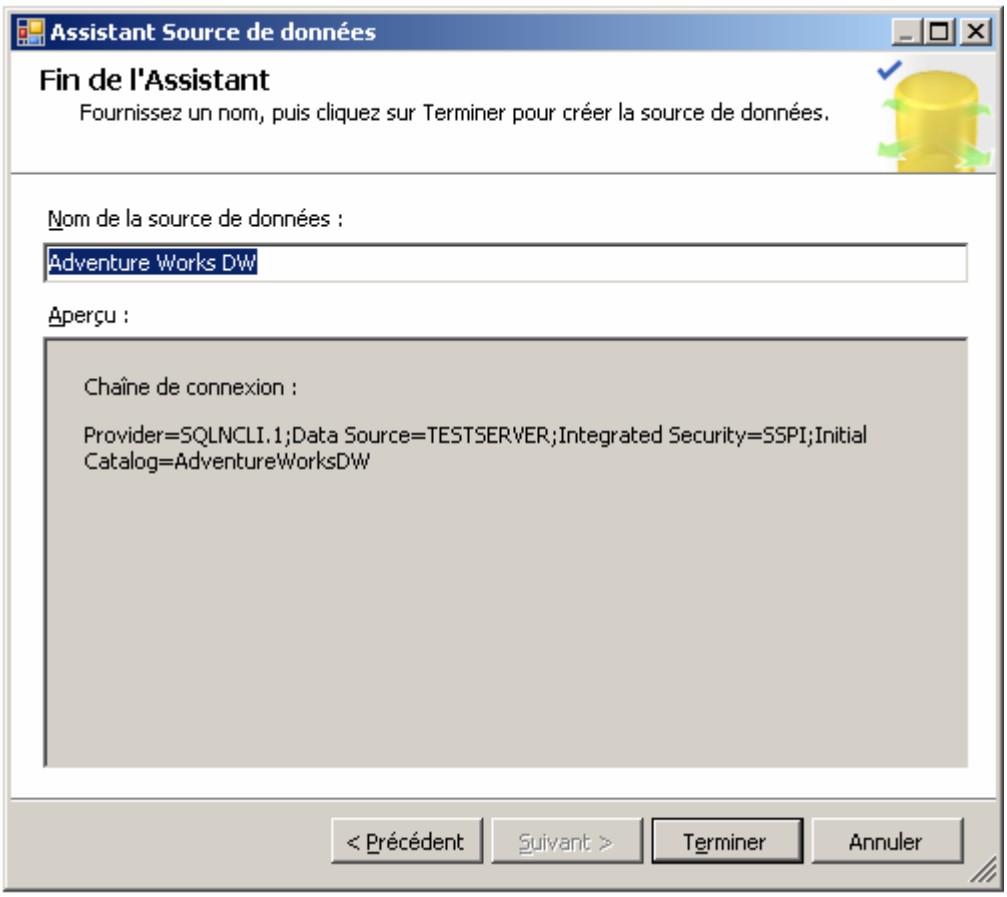


Figure 3-10 : Assistant Source de données - Fin de l'Assistant

L'explorateur de solutions possède maintenant une source de données nommée « Adventure Works DW » (Figure 3-11).

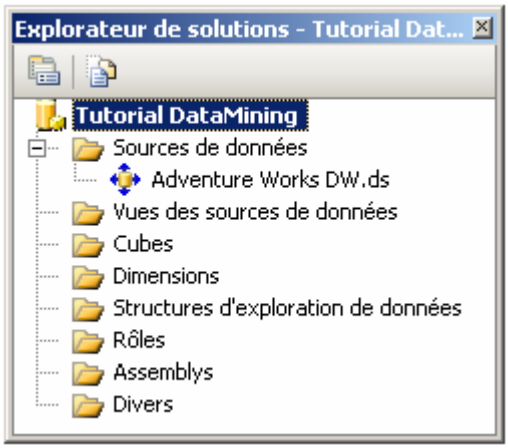


Figure 3-11 : Fenêtre « Explorateur de solutions » - Avec Source de données

Une fois une Source de données configurée, l'étape suivante consiste à créer des « Vues des sources de données ».

3.2 Créer une Vues des sources de données

Une « Vues de sources de données » peut être apparentée à un schéma de base de données. C'est dans une « Vue de base de données » que nous spécifions les tables/vues de la base de données que nous voulons utiliser dans notre solution SSAS.

Comme pour la création des « Sources de données », il suffit de faire un clic droit sur le dossier « Vues des sources de données » et de choisir l'option « Nouvelle vue de sources de données… » (Figure 3-12).

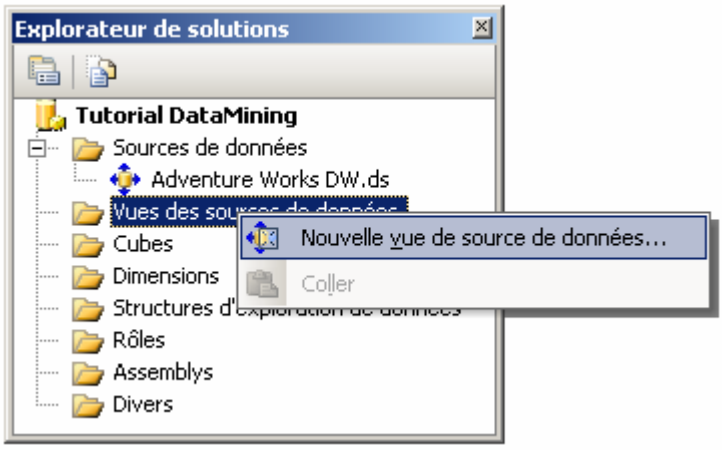


Figure 3-12 : Fenêtre « Explorateur de solutions » - Nouvelle vue de source de données

L'assistant « Vue de source de données » (Figure 3-13) apparaît à l'écran. Cliquer sur « Suivant ».

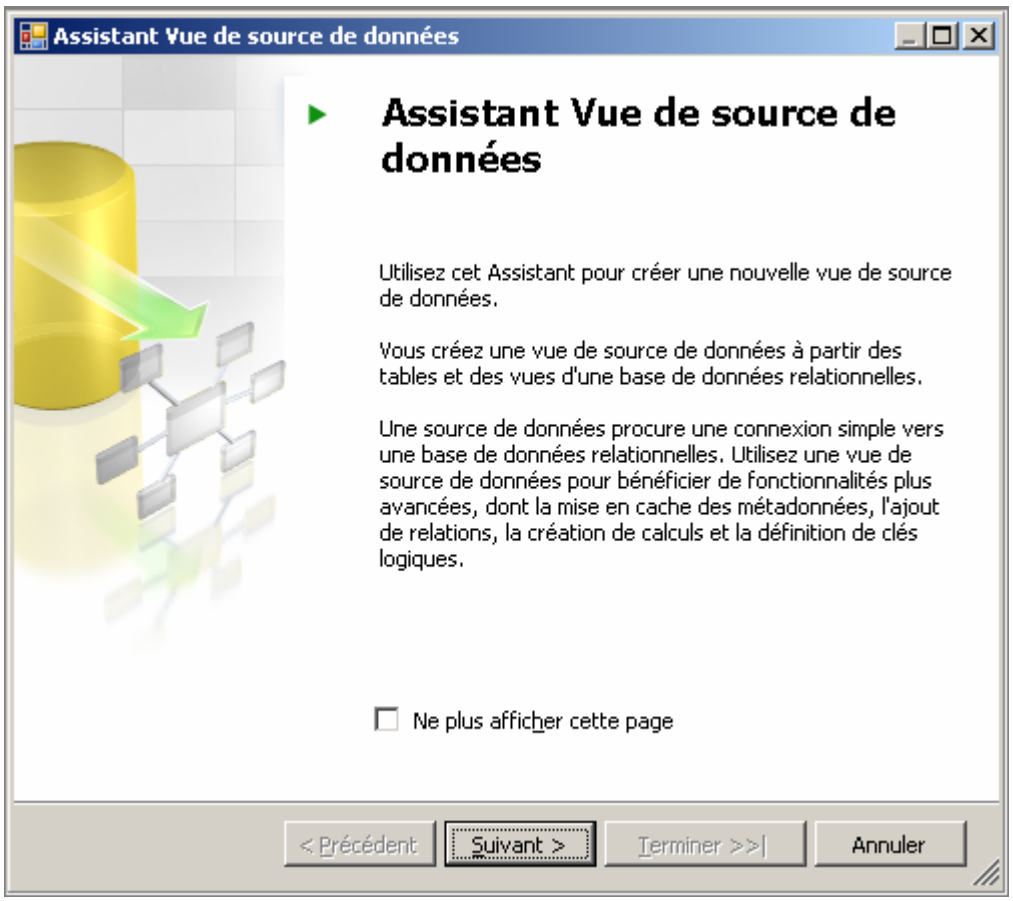


Figure 3-13 : Assistant Vue de source de données

L'écran « Sélectionner une source de données » (Figure 3-14) permet, comme son nom l'indique, de choisir une source de donnée.
Le Panneau de gauche affiche toutes les sources de données définies à l'étape «

Créer une Sources de données ». Etant donné que nous n'avons configuré qu'une seule source de données, celle-ci est sélectionnée par défaut.
L'écran « Sélectionner une source de données » propose de créer, si ce n'est pas encore fait, une « Nouvelle source de données… », via le bouton correspondant.

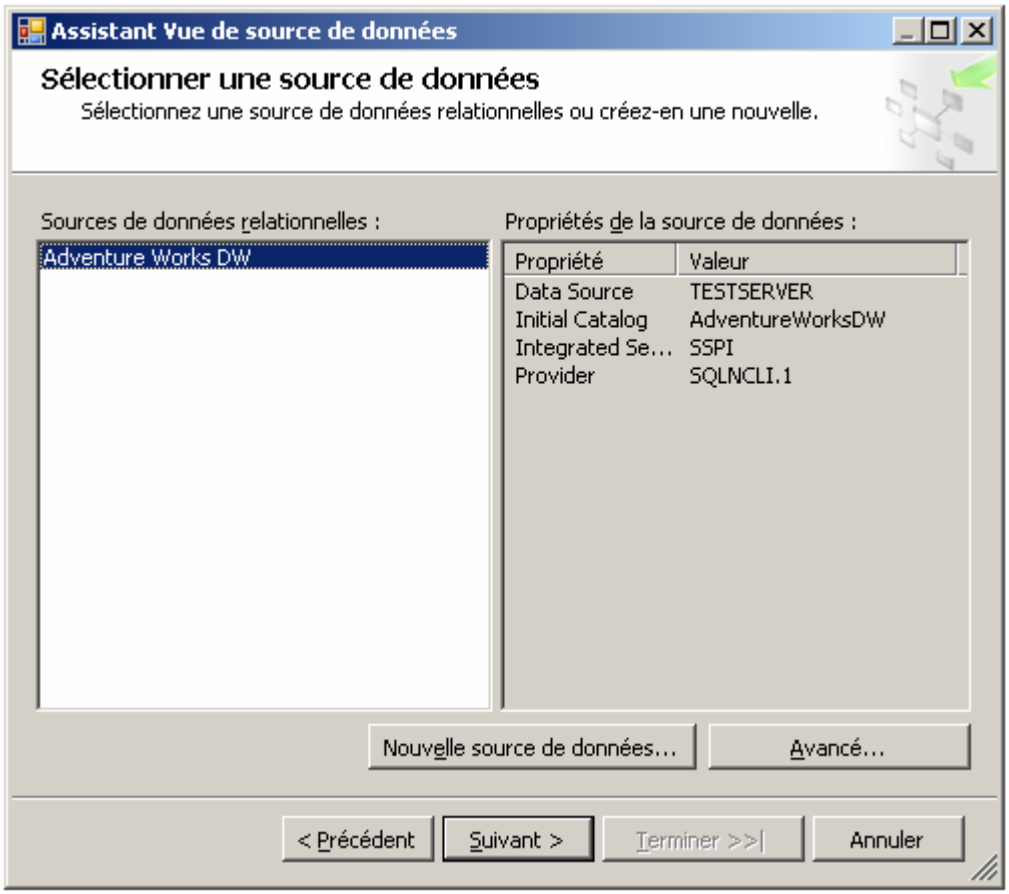


Figure 3-14 : Assistant Vue de source de données - Sélectionner une source de données.

Le bouton « Avancé… » permet, dans la source de données sélectionnée, de choisir le schéma de la base de données ainsi que d'extraire les différentes relations entre les tables de la base de données (Figure 3-15).

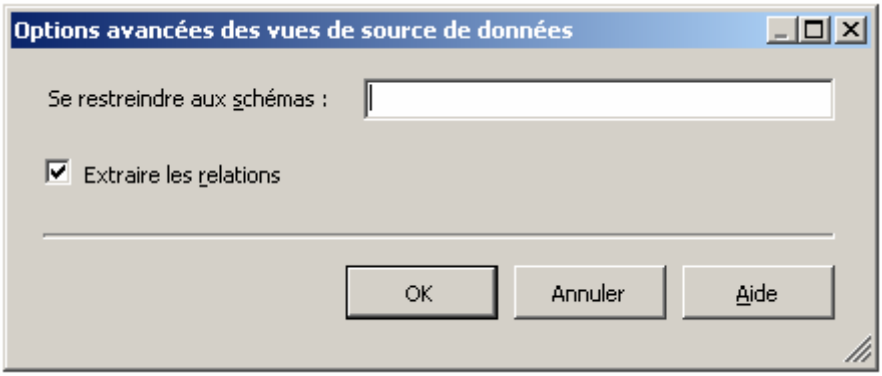


Figure 3-15 : Assistant Vue de source de données - Options avancées des vues de source de données

Après avoir choisi la source de données qui nous permettra de choisir les tables/vues de la base de données, cliquer sur « Suivant »

A l'écran « Sélectionner des tables et des vues » (Figure 3-16), nous sélectionnons la vue « dbo.vTargetMail » soit en double cliquant sur son nom dans le volet de gauche « Objets disponibles : », soit en la choisissant et en cliquant sur le bouton « > ».

Répéter la même opération afin d'ajouter la table « dbo.ProspectiveBuyer »

Nb. La vue « vTargetMail » contient le détail d'une liste une d'acheteurs qui ont ou non acheté un vélo. Les acheteurs de vélos sont flagés à 1 dans le champ [BikeBuyer].
La table « ProspectiveBuyer » sera la table de client potentiel à qui s'adressera cette compagne de publipostage.

Une fois la vue et la table ajoutées dans le volet de droite « Objets inclus : », cliquer sur « Suivant ».

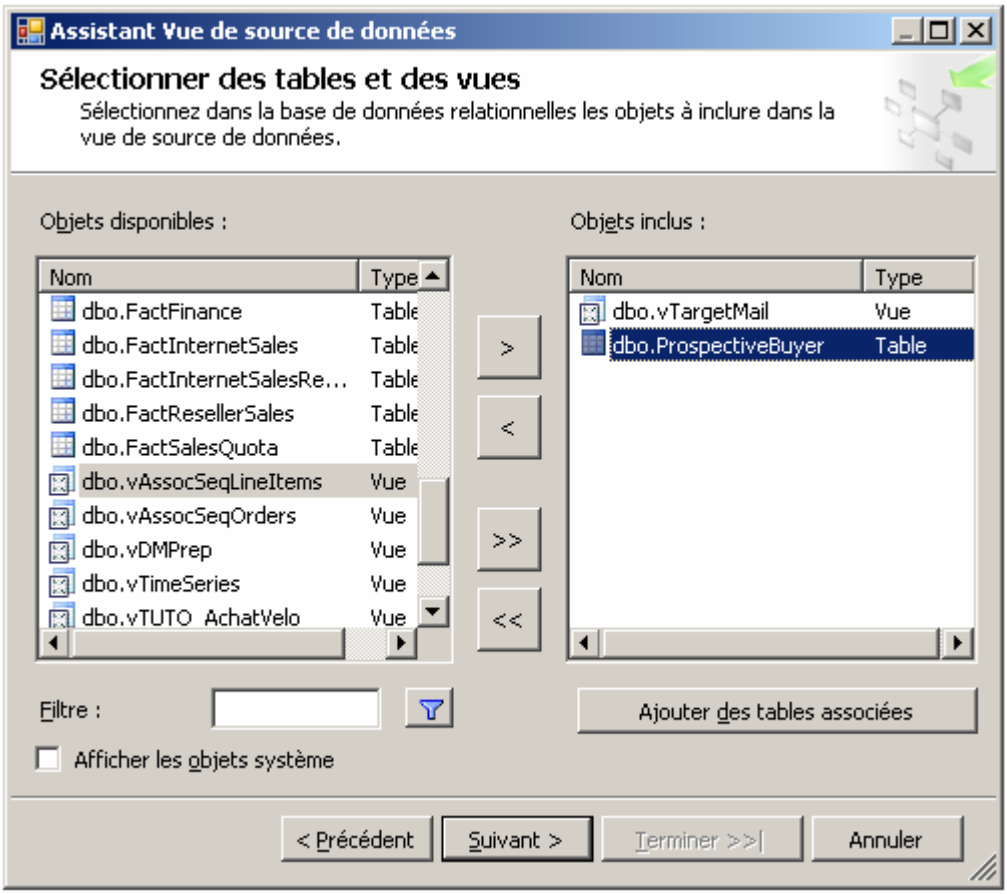


Figure 3-16 : Assistant Vue de source de données - Sélectionner des tables et des vues

Afin d'achever la configuration de la « Vue de source de données », il ne reste plus qu'à lui donner un nom (ici « Adventure Work DW ») et à cliquer sur « Terminer » (Figure 3-17).

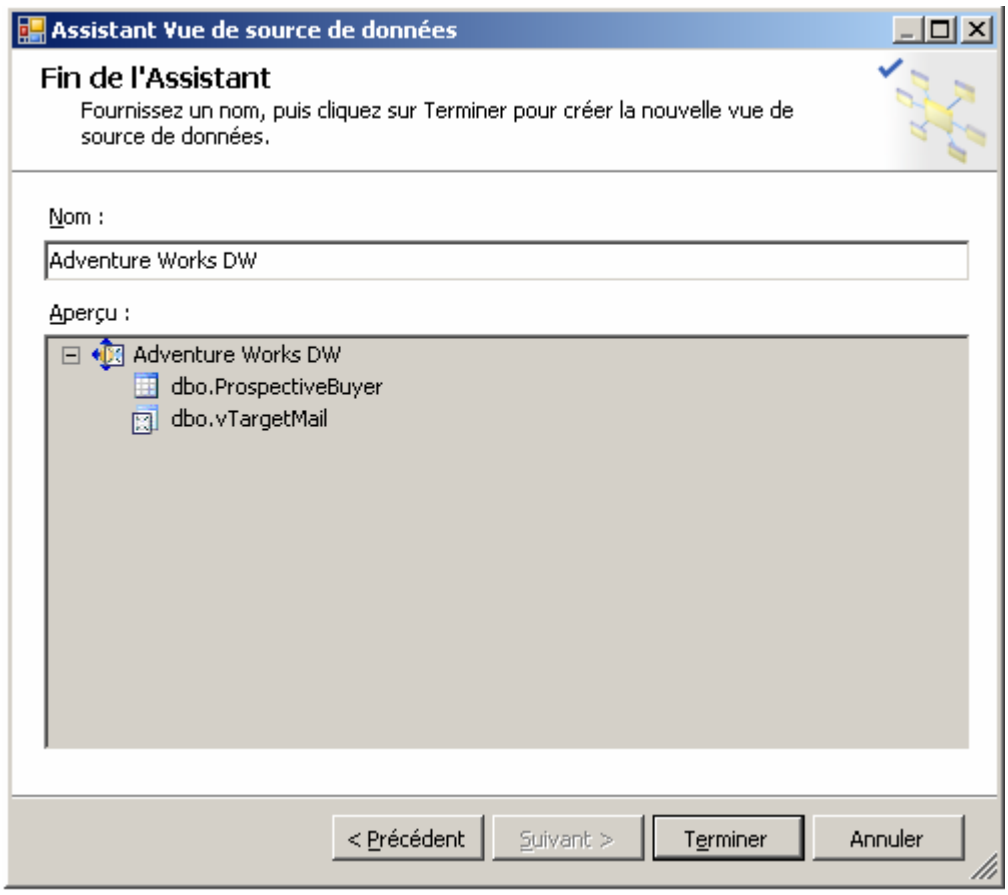


Figure 3-17 : Assistant Vue de source de données - Fin de l'Assistant

Désormais, notre volet « Explorateur de solutions » s'est enrichi d'un nouvel élément (Figure 3-18).

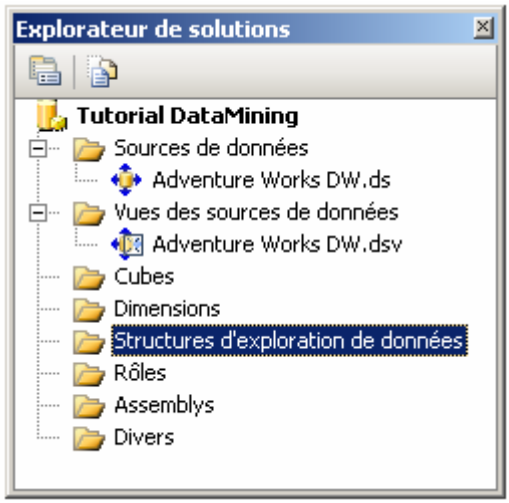


Figure 3-18 : Fenêtre « Explorateur de solutions »- Avec une Vue des sources de données

3.3 Créer une Structures d'exploration de données

Aux étapes précédentes, nous n'avons fait « que » préparer l'accès aux données.
Pour information, lors d'un projet ayant pour but la création d'un cube OLAP, ces étapes doivent aussi être réalisées.

A partir de ce point, nous entrons dans le monde du Data Mining.

Pour bien commencer, définissons le terme de « Structures d'exploration de données » :
Une « Structure d'exploration de données » définit quelles sont les données que nous utilisons, quel(s) est/sont le/les algorithme/s que nous utilisons pour l'analyse, quels sont les types de données et les types de contenu des données sources, etc…

Afin de créer notre première « Structure d'exploration de données », nous utilisons l'assistant d'exploration de données.

Nous commençons par un clic droit sur l'onglet « Structure d'exploration de données » de l'« Explorateur de solutions » et nous choisissons dans le menu contextuel « Nouvelle structure d'exploration de données… » (Figure 3-19).

Une « Structure d'exploration de données » peut être aussi appelée « Modèle d'exploration de données ».

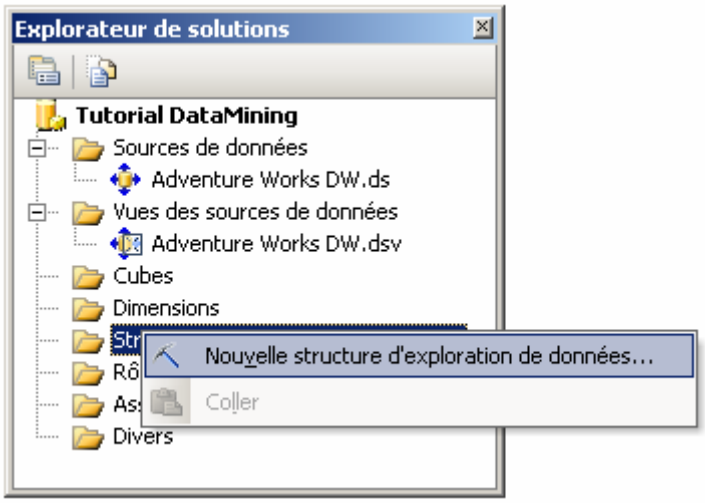


Figure 3-19 : Fenêtre « Explorateur de solutions » - Nouvelle structure d'exploration de données

L'écran « Assistant Exploration de données » (Figure 3-20) s'affiche. Cliquer sur « Suivant » pour continuer.

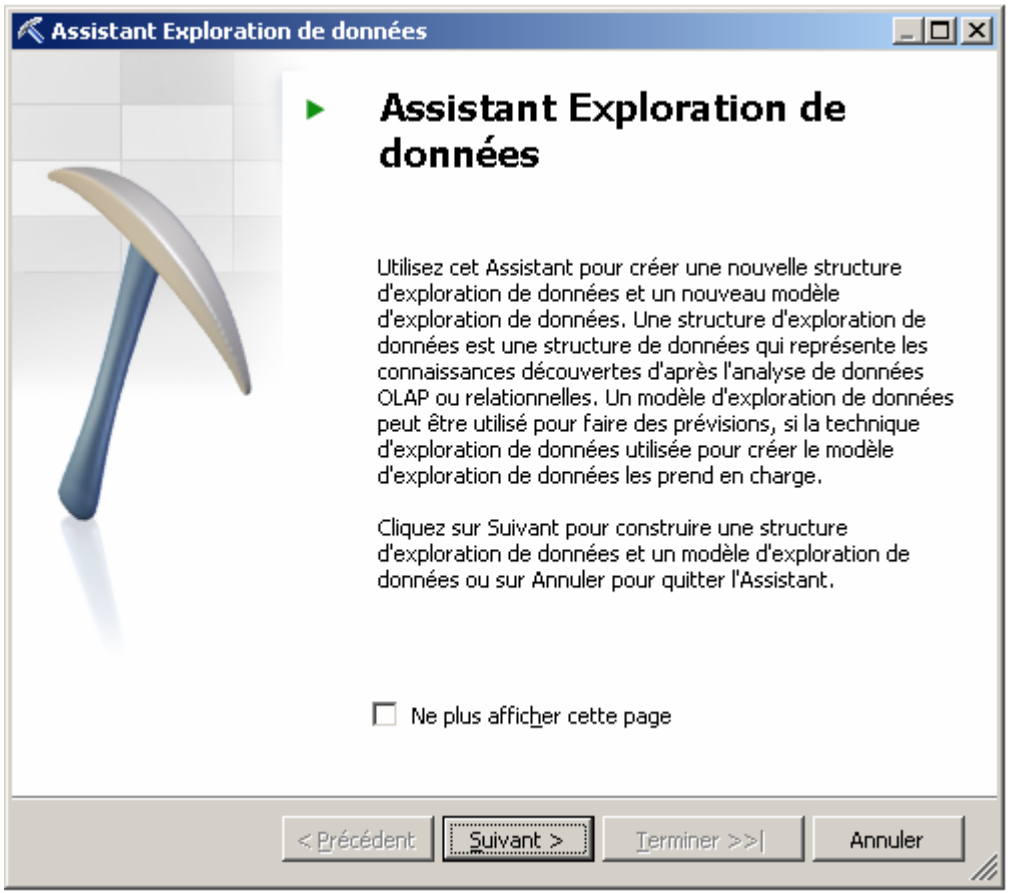


Figure 3-20 : Assistant Exploration de données

A l'écran « Sélectionner la méthode de définition », vous avez la possibilité de choisir entre deux options pour la méthode à utiliser lors de la création de la structure d'exploration de données :

- A partir d'une base de données relationnelles ou d'un entrepôt de données qui existe déjà ;
- A partir d'un cube existant.

La première option signifie que nous allons puiser les données sources directement dans des tables.
La deuxième option signifie que nous allons puiser les données sources dans un cube OLAP.

Dans le cadre de ce tutorial, nous choisissons la première option (Figure 3-21) et nous cliquons sur « Suivant ».

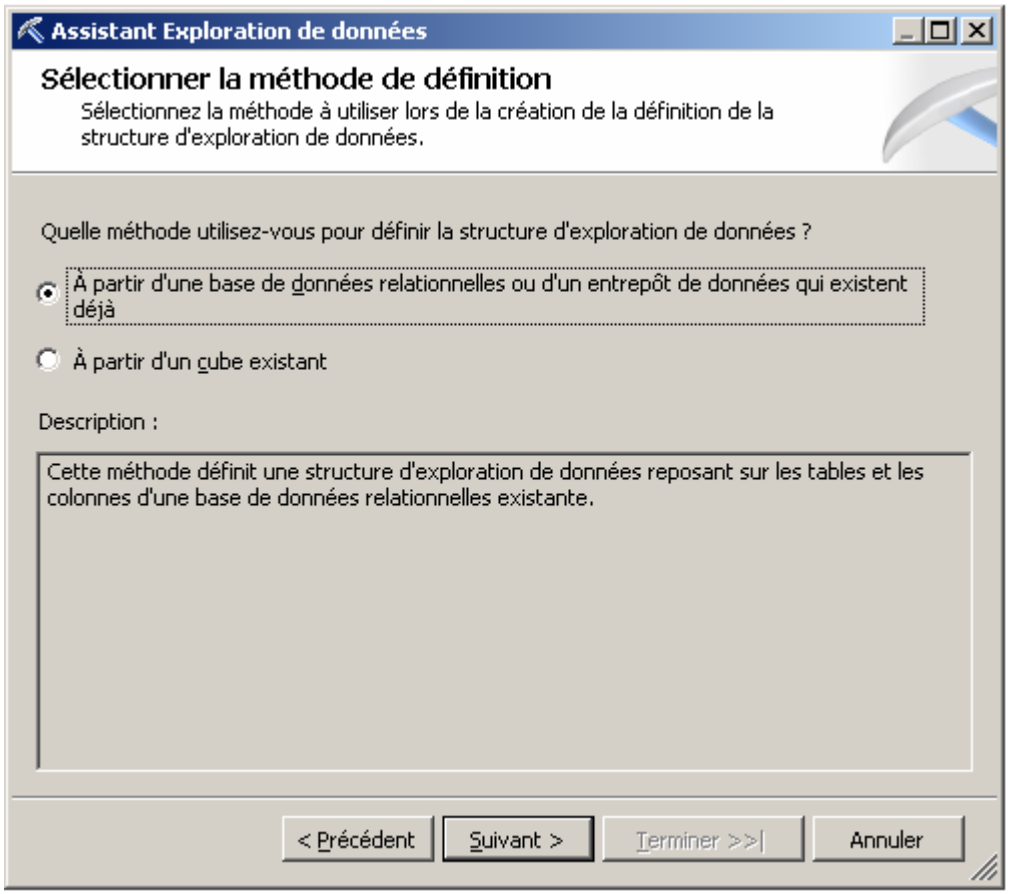


Figure 3-21 : Assistant Exploration de données - Sélectionner la méthode de définition

Ä l'écran « Sélectionner la technique d'exploration des données » (Figure 3-22), nous choisissons l'algorithme « Algorithme MDT (Microsoft Decision Trees » de la liste déroulante.

Plusieurs algorithmes d'exploration de données sont disponibles dans VS 2005 – BI. Ceux-ci sont expliqués plus en détail dans mon rapport final.
En voici un petit rappel de ceux-ci :

- Algorithme MAR (Microsoft Association Rules) ;
- Algorithme MDT (Microsoft Decision Trees) ;
- Algorithme MNB (Microsoft Naive Bayes) ;
- Algorithme MNN (Microsoft Neural Network) ;
- Algorithme MSC (Microsoft Sequence Clustering) ;
- Algorithme MTS (Microsoft Time Series) ;
- Clusters Microsoft ;
- MLR (Microsoft Linear Regression) ;
- MLR (Microsoft Logisitc Regression).

Durant ce tutorial, nous utilisons deux algorithmes supplémentaires en plus de l'« Algorithme MDT » : l'« Algorithme MNB » et le « Clusters Microsoft ». Ils seront configurés plus tard dans ce tutorial.

Après avoir choisi l'« Algorithme MDT », cliquer sur « Suivant »

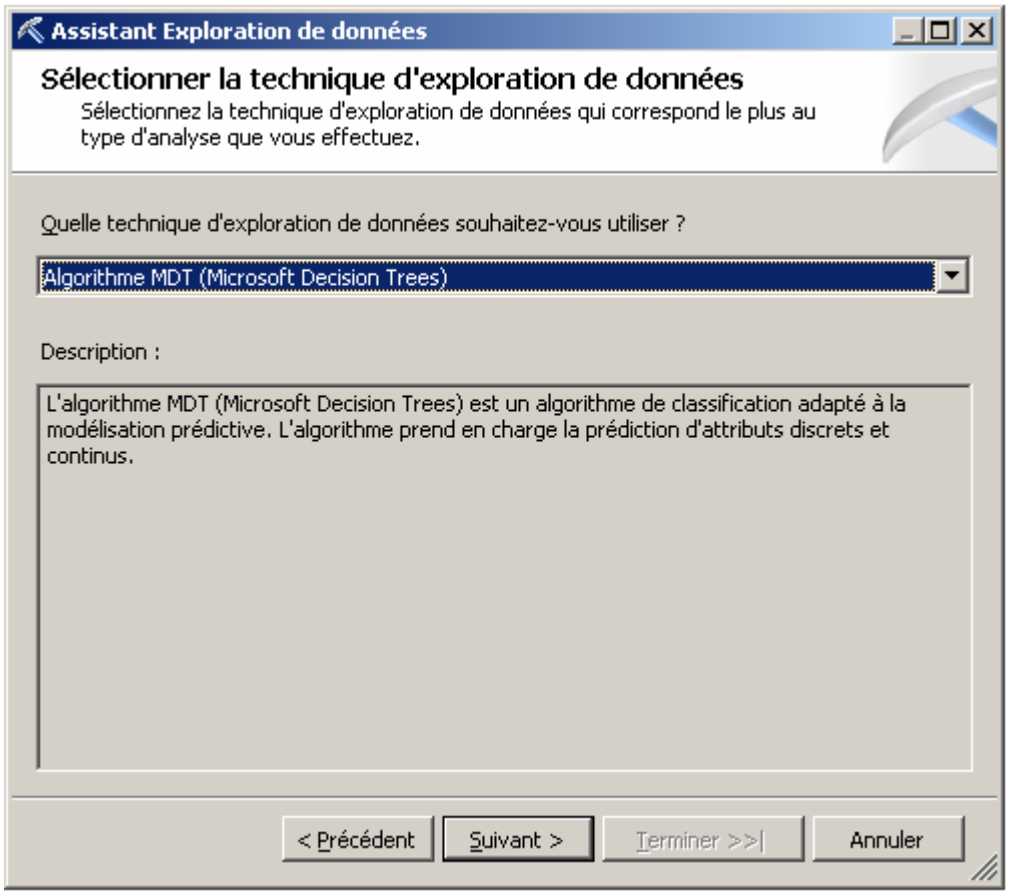


Figure 3-22 : Assistant Exploration de données - Sélectionner la technique d'exploration de données

A l'écran « Sélectionner une source de données » (Figure 3-23), le volet de gauche nous présente toutes les « Vues de sources de données » disponibles dans notre solution de Data Mining. Quant au volet de droite, celui-ci nous affiche, selon la « Vue de sources de données » sélectionnée, les diverses tables disponibles dans la vue.

Etant donné que nous avons configuré une seule « Vue de sources de données », celle-ci est automatiquement sélectionnée. Nous continuons en cliquant sur « Suivant ».

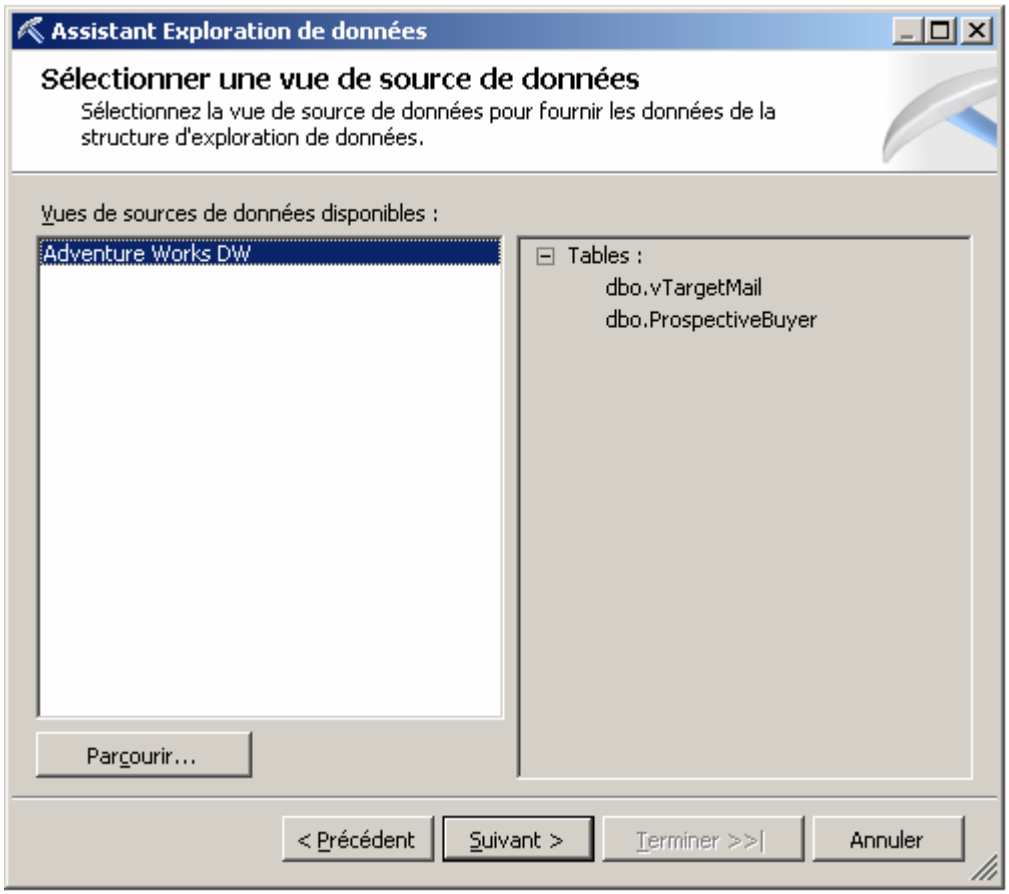


Figure 3-23 : Assistant Exploration de données - Sélectionner une vue de source de données

A l'écran « Spécifier les types des tables » (Figure 3-24), nous marquons la case à cocher de la colonne « Cas » en visu de la vue « vTargetMail » et nous cliquons sur « Suivant ».

La colonne « Imbriqué » signifie que nous vonlons ajouter à la table des « Cas » une seconde table.

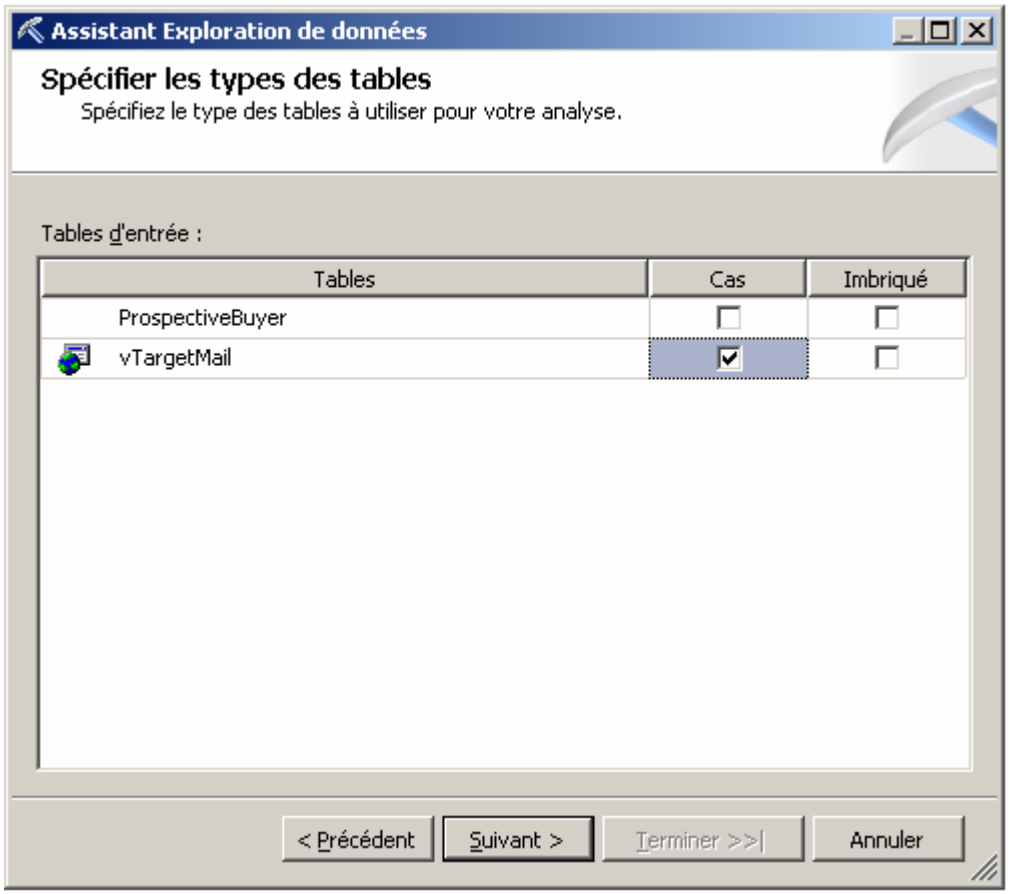


Figure 3-24 : Assistant Exploration de données - Spécifier les types de tables

A l'écran « Spécifier les données d'apprentissage » (Figure 3-25), il faut marquons la case à cocher de la colonne « Clé » de champ « CustomerKey ».
En cochant cette case, nous indiquons à l'algorithme quelle est la clé unique des enregistrements des données sources.
Il est nécessaire aussi de cocher la case à cocher « Prévisible » du champ « BikeBuyer » qui indique à l'algorithme que c'est ce champ qui doit être prédit.
Une fois que nous avons indiqué à l'algorithme quel est le champ clé et le/les champ(s) à prédire, le bouton « Suggérer » devient alors actif..
En cliquant sur ce bouton, l'algorithme exécute quelques traitements afin de proposer à l'utilisateur les colonnes d'entrée qu'il juge pertinentes. Nous cliquons sur ce bouton.

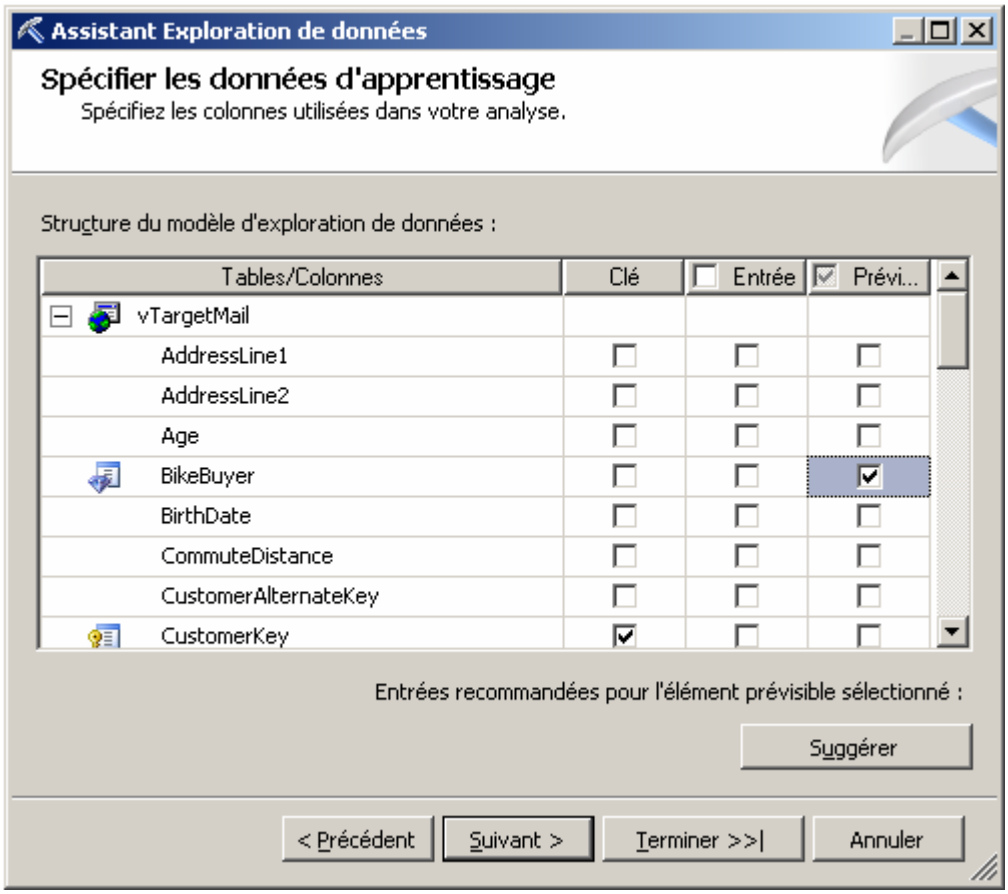


Figure 3-25 : Assistant Exploration de données – Spécifier les données d'apprentissage

L'écran « Suggérer des colonnes associées » (Figure 3-26) s'affiche. Par défaut, en regard de chaque champ qui propose un coefficient de corrélation supérieur à 0.05, une coche dans la colonne « Entrée » est posée.
La colonne âge est automatiquement cochée (coefficient de 0.068) et nous ajoutons des colonnes supplémentaires pour déterminer les données d'apprentissage :

- CommuteDistance distance entre le travail et la maison
- EnglishEducation formation scolaire (en anglais)
- EnglishOccupation profession (en anglais)
- FirstName prénom
- Gender sexe
- GeographyKey point géographique
- HouseOwnerFlag propriétaire de la maison
- LastName nom
- MaritalStatus état civil
- NumberCasOwned nombre de voiture
- NumberChildrenAtHome nombre d'enfants à la maison
- Region région d'habitation
- TotalChildren nombre d'enfants total
- YearlyIncome salaire annuel

Une fois les colonnes choisies, cliquer sur « OK ».

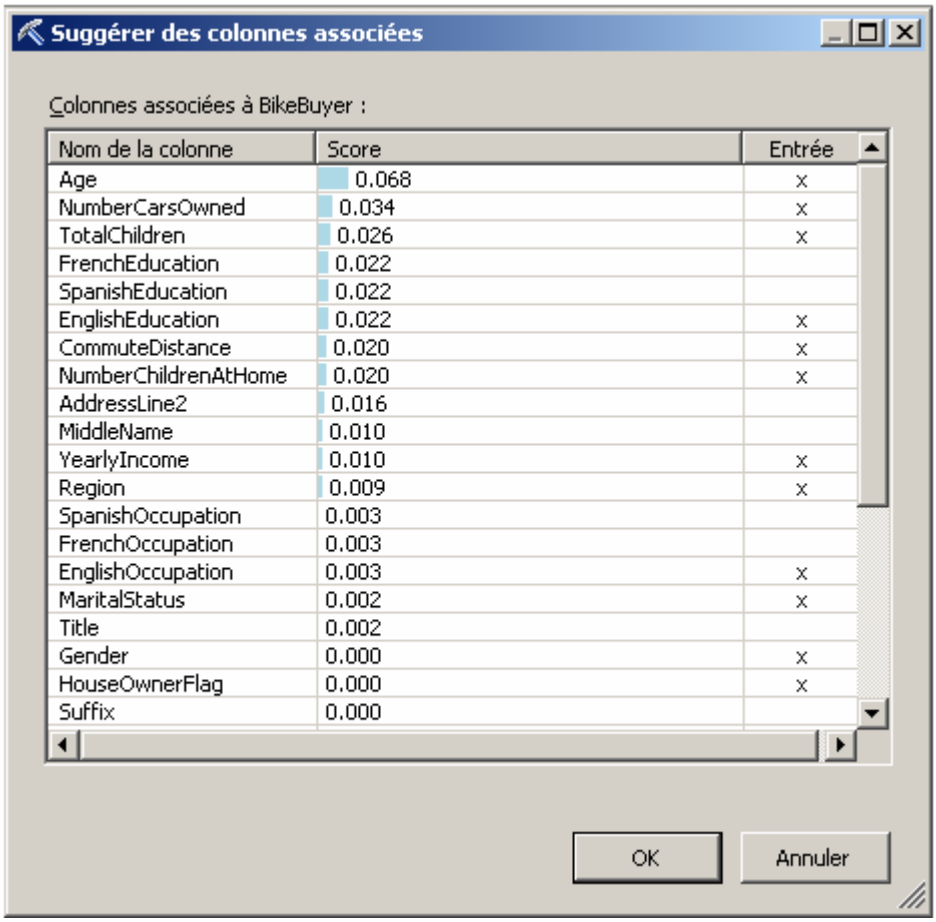


Figure 3-26 : Assistant Exploration de données – Suggérer des colonnes associées

De retour à l'écran « Spécifier les données d'apprentissage » (Figure 3-25), nous contrôlons que tous les champs spécifiés à l'écran « Suggérer des colonnes associées » (Figure 3-26) sont cochés dans la colonne « Entrée ». Une fois le contrôle effectué, nous cliquons sur « Suivant ».

L'écran « Spécifier le type de contenu et de données des colonnes » (Figure 3-27) s'affiche. Dans cet écran nous devons spécifier quel est le type de contenu (Discrete/Continuous/etc…) et le type de données (Text/Long/etc…) des champs choisis.

Les types de contenu et les types de données doivent être les suivants (Tableau 3-1 : Type de contenu et type de données) :

	Type de contenu	Type de donnée
Age	Discretized	Long
Bike Buyer	Discrete	Long
Commute Distance	Discrete	Text
Customer Key	Key	Long
English Education	Discrete	Text
English Occupation	Discrete	Text
First Name	Discrete	Text
Gender	Discrete	Text
Geography Key	Discrete	Text
House Owner Flag	Discrete	Long
Last Name	Discrete	Text
Marital Status	Discrete	Text
Number Cas Owned	Discrete	Long
Number Children At Home	Discrete	Long
Region	Discrete	Text
Total Children	Discrete	Long

	Type de contenu	Type de donnée
Yearly Income	Continuous	Double

Tableau 3-1 : Type de contenu et type de données

Une fois les types de contenu et les types de données paramétrés, cliquer sur « Suivant »

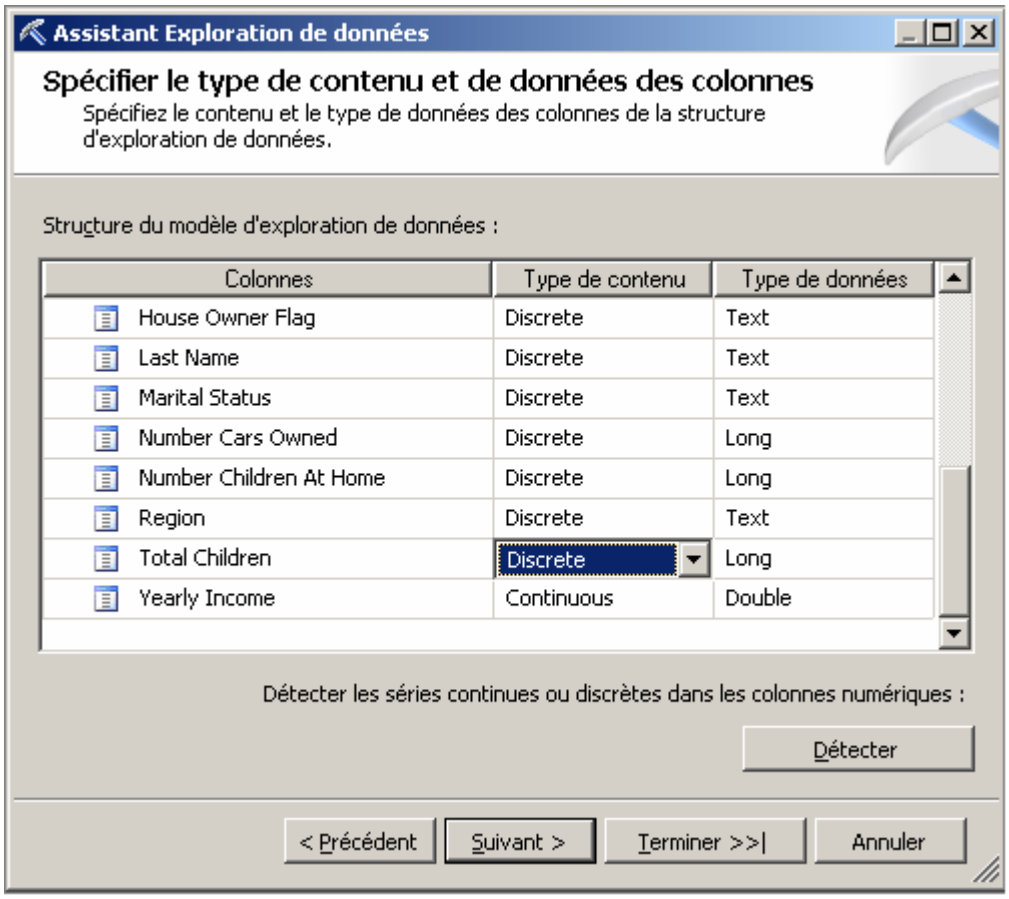


Figure 3-27 : Assistant Exploration de données - Spécifier le type de contenu et de données des colonnes

Pour mettre fin à l'assistant, à l'écran « Fin de l'assistant » (Figure 3-28), il nous reste plus qu'à donner un nom à la structure d'exploration de données créé. Nous la nommons « Publipostage ».

Il faut également indiqué un nom au modèle d'exploration de données. Nous le nommons MDT.

Pour ce type de Data Mining prévisionnel, nous cochons aussi la case à cocher « Accepter l'extraction », qui nous permet, dans une prochaine étape, de consulter les données d'apprentissage du modèle d'exploration des données.

Pour finir, cliquer sur « Terminer ».

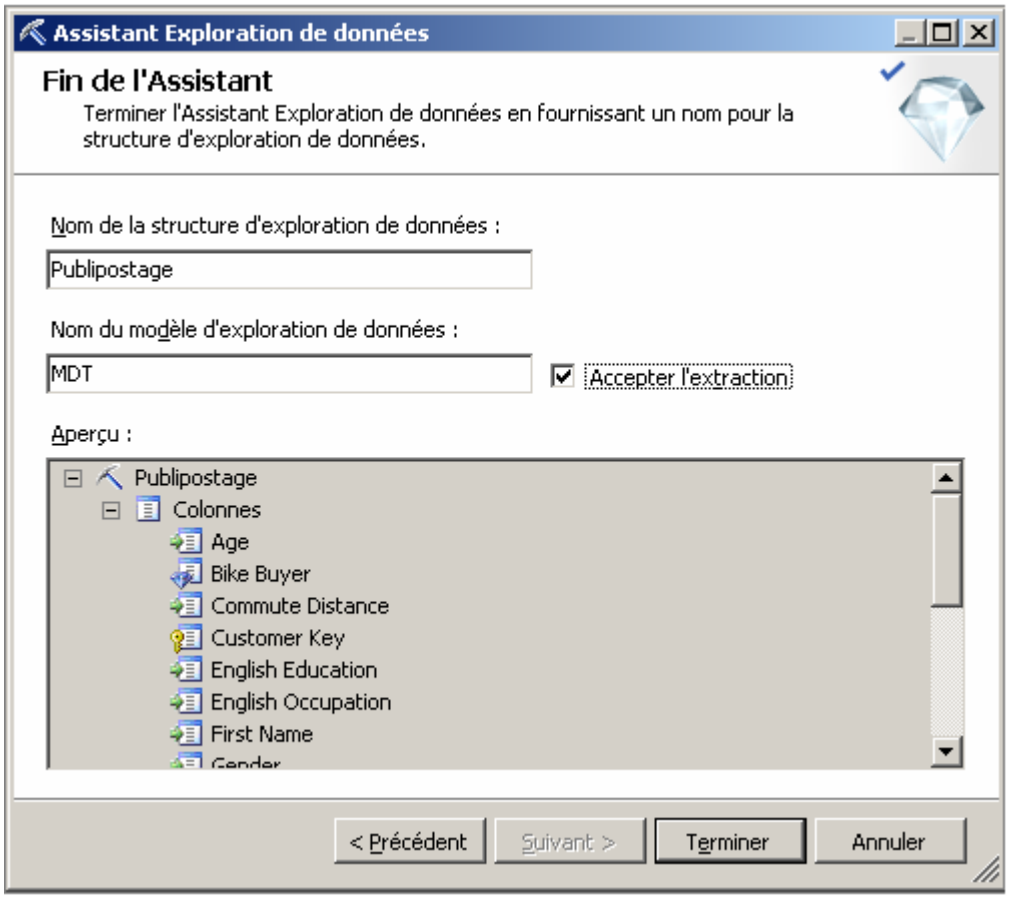


Figure 3-28 : Assistant Exploration de données - Fin de l'assistant

L'« Explorateur de solutions » (Figure 3-29) s'est enrichi d'un nouvel objet : « Publipostage.dmm ».

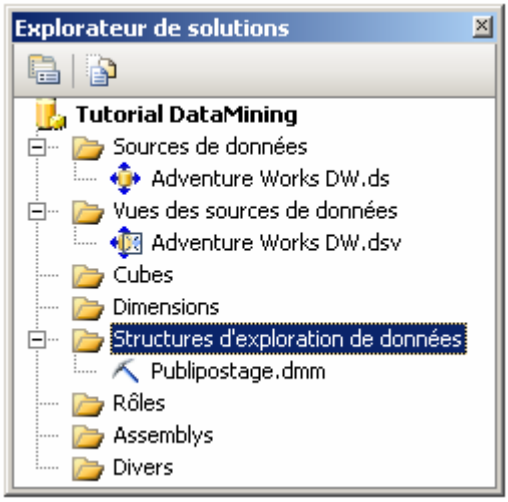


Figure 3-29 : Fenêtre « Explorateur de solutions » - Avec une Structure d'exploration de données

3.4 Navigation dans une Structures d'exploration de données – MDT

Jusqu'à maintenant, nous avons préparé la source de données, la table de données ainsi qu'un model d'exploration de données.

Dès à présent nous allons apprendre à « Explorer » une « Structure d'exploration de données ».

Au chapitre précédent, nous avons utilisé l'algorithme « Microsoft Decision Tree » pour créer un modèle d'exploration de données. Maintenant nous allons l'interpréter.

Pour l'ouvrir, il suffit de double cliquer sur « Publipostage.dmm » ou alors de faire un clic droit et de choisir « Ouvrir » dans le menu contextuel (Figure 3-30).

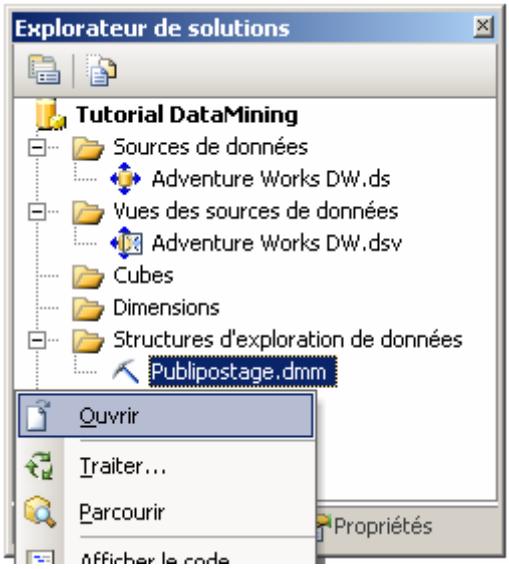


Figure 3-30 : Fenêtre « Explorateur de solutions » - Ouvrir… MDT

A la suite de l'ouverture de « Publipostage.dmm », VS 2005 affiche à l'écran de nouveaux onglets (Figure 3-31) :

- Structure d'exploration de données ;
- Modèles d'exploration de données ;
- Visionneuse de modèle d'exploration de données ;
- Graphique d'analyse de précision ;
- Prévision de modèle d'exploration de données.

Nous allons passer à travers chacun de ces onglets afin de les expliquer.

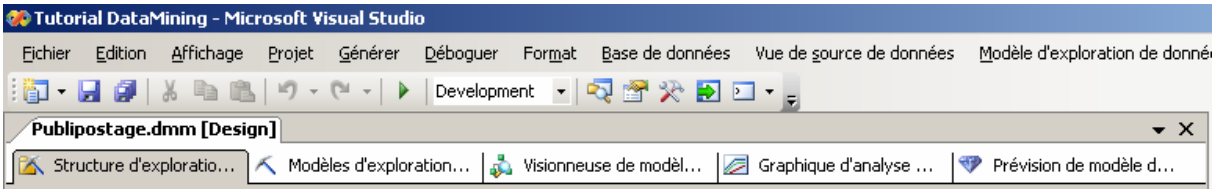


Figure 3-31 : Fenêtre « Publipostage.dmm »

3.4.1 Structure d'exploration de données

Cet onglet (Figure 3-32) permet d'ajouter des champs dans le model d'exploration de données et il permet aussi de redéfinir les types de données ainsi que les types de contenu des champs de la « Structure d'exploration de données ».

Pour ajouter un champ, il suffit, à partir de la « Vue de source de données » au centre de l'écran (ici « vTargetMail ») de choisir le champs désiré en cliquant dessus et, tout en le maintenant, le glisser sur la structure d'exploration de données du volet de gauche. Il est aussi possible d'ajouter un champ, plus simplement, via l'icône .

Ensuite, il est nécessaire de choisir son type de données ainsi que son type de contenu en cliquant sur ce nouveau champ (dans le volet de gauche) et, dans le volet « Propriétés, Type de données », de choisir son type de contenu (« Content ») et son type de données (« Type »). Pour la modification d'un type de contenu ou d'un type de données d'un champ faisant déjà partie de la structure d'exploration de données, la méthode est la même.

A partir de cet onglet, il est aussi possible d'ajouter une table dérivée en cliquant sur l'icône .

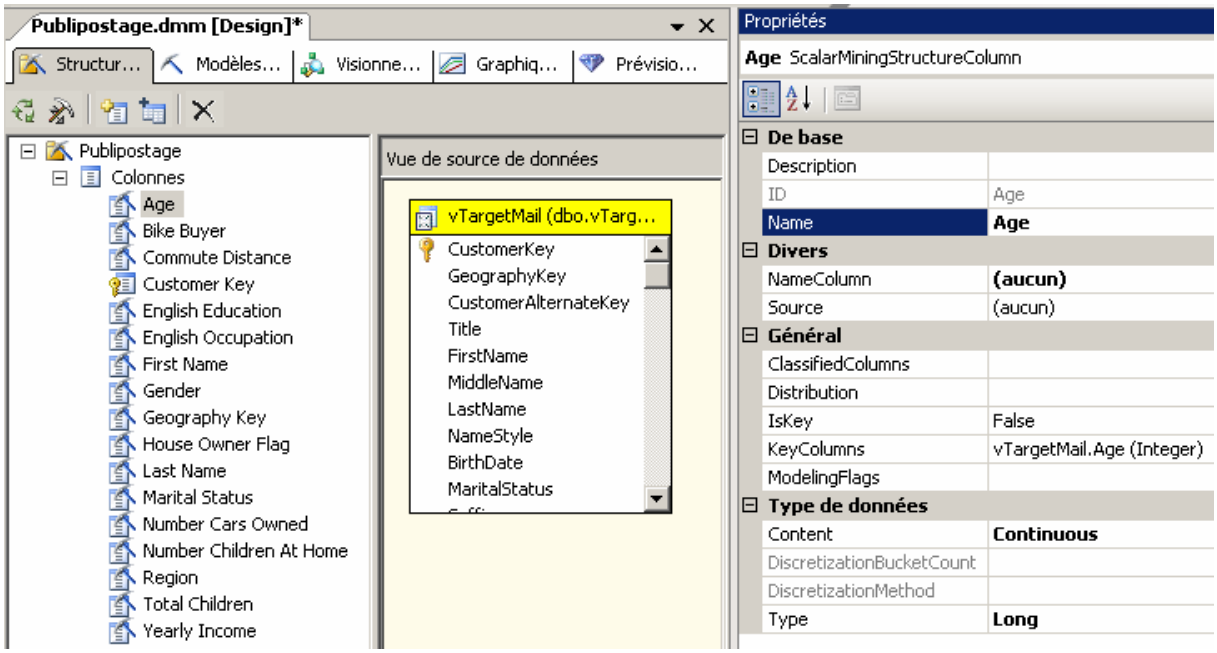


Figure 3-32 : Fenêtre « Publipostage.dmm » - Onglet « Structure d'exploration de données »

3.4.2 Modèles d'exploration de données

Dans cet onglet (Figure 3-33), il est encore possible de modifier les types de contenu ainsi que les type de données des champs de la structure d'exploration de données.

C'est aussi dans cet onglet, qu'il est possible d'ajouter un nouveau modèle d'exploration de données (icône).

Comme il est possible, dans cet onglet, de modifier quel est le comportement des champs pour l'algorithme sélectionné, voici les différents comportements possibles disponibles dans les listes déroulantes (Figure 3-34) en vis-à-vis des champs :

- Ignorer Informe l'algorithme qu'il ne doit pas utiliser ce champ au moment de la génération du modèle ;
- Input Informe l'algorithme qu'il doit utiliser ce champ au moment de la génération du modèle ;
- Predict Informe l'algorithme que ce champ est un champ à prédire ;
- PredictOnly Informe l'algorithme que ce champ est un champ à prédire.

La différence entre « Predict » et « PredictOnly » est que, dans le cas de « Predict », si le modèle doit prédire plusieurs champs, l'algorithme peut utiliser ce champs en tant que champ d'entrée pour la prévision du deuxième champ et inversement.
Le fait de choisir « PredictOnly » informe l'algorithme que le champ est à ignorer lors de la génération de modèle du deuxième champ et inversement.

Nb. S'il y a un seul champ à prédire, il peut être soit « Predic » ou « PredicOnly ».

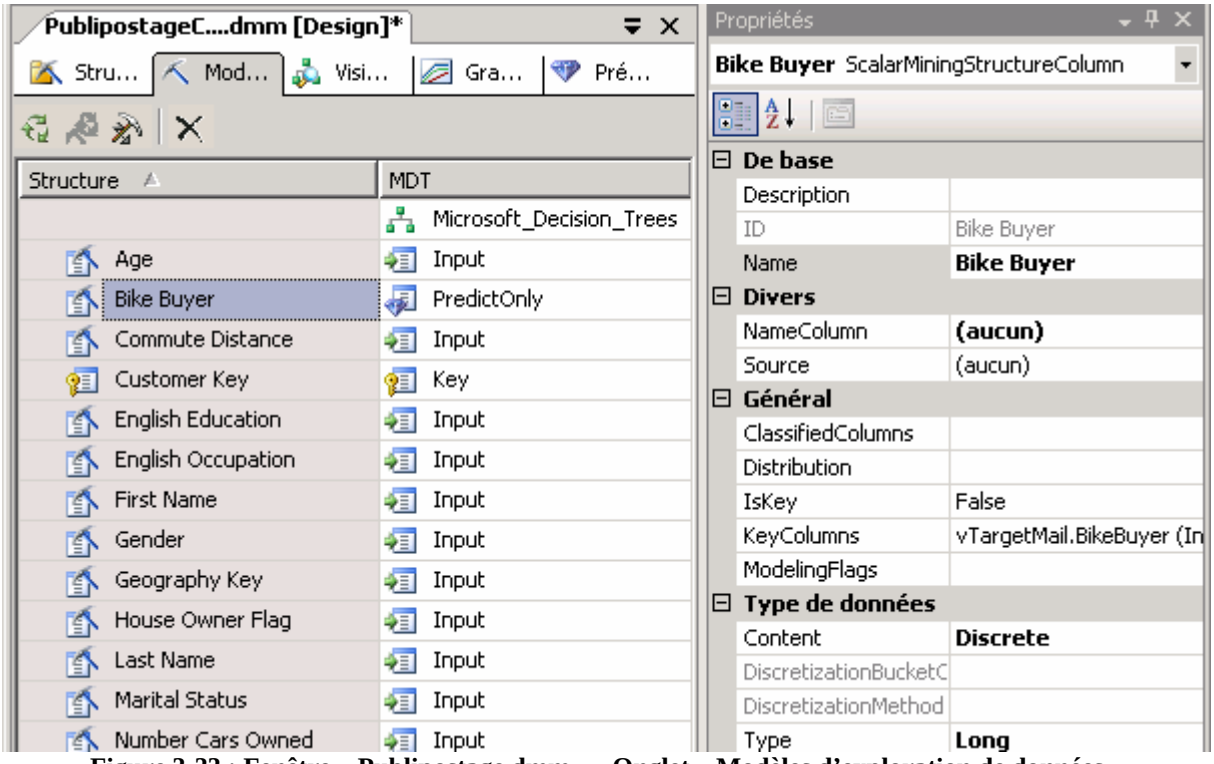


Figure 3-33 : Fenêtre « Publipostage.dmm » - Onglet « Modèles d'exploration de données » Algorithme MDT

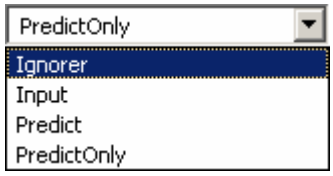


Figure 3-34 : Fenêtre « Publipostage.dmm » - Onglet « Modèles d'exploration de données » - Liste déroulante « choix du comportement »

3.4.3 Visionneuse d'exploration de données

La « Visionneuse d'exploration de données » permet d'explorer le modèle d'exploration de données.

Lors du clic sur cet onglet, VS 2005 – BI effectue une vérification du contenu sur lequel le modèle d'exploration de données s'appuie (Figure 3-35).
Etant donné que nous n'avons pas encore traité la structure de données, VS 2005 – BI nous propose également de traiter le modèle d'exploration de données (Figure 3-36).

A chacune des questions de VS 2005 – BI, nous répondons par l'affirmative.

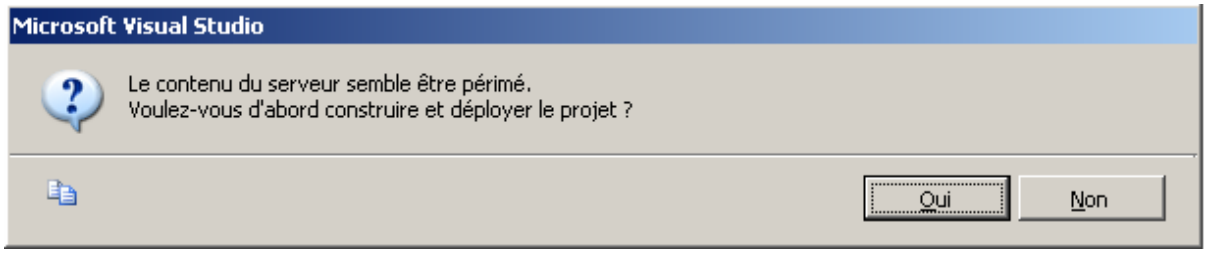


Figure 3-35 : Fenêtre « Publipostage.dmm » - Onglet « Visionneuse d'exploration de données » - Construction et déploiement du projet

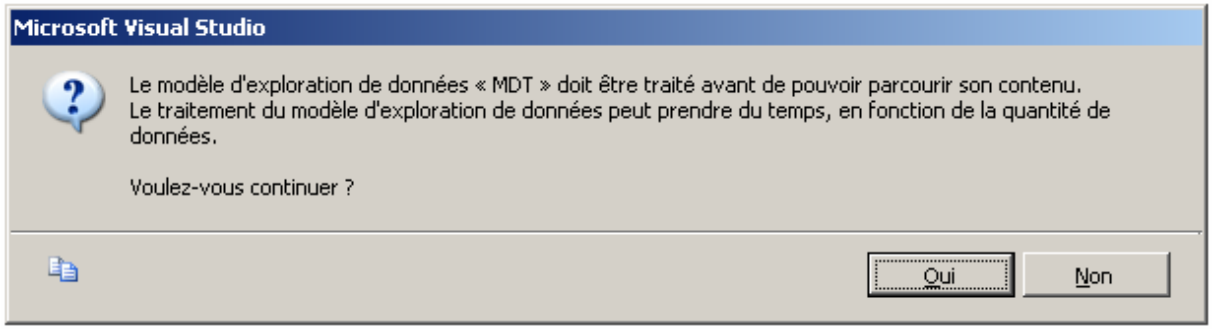


Figure 3-36 : Fenêtre « Publipostage.dmm » - Onglet « Visionneuse d'exploration de données » - Traitement du modèle d'exploration de données

A la suite de la deuxième question, l'écran « Traiter Modèle d'exploration de données » (Figure 3-37) s'affiche.

Dans cet écran, il est possible de choisir, si nous en avons créés plusieurs, quels sont les modèles de données que nous désirons mettre à jour en supprimant via le bouton « Supprimer » les modèles que nous ne désirons pas mettre à jour.

Cet écran permet aussi d'analyser l'impact sur les objets dépendants de ce modèle via le bouton « Analyse d'impact ».

Le bouton « Modifier les paramètres » permet d'afficher un écran contenant tous les paramètres de l'algorithme. Ces paramètres étant différents d'un algorithme à un autre, je vous laisse vous référer au rapport final afin d'en prendre connaissance.

Pour traiter notre modèle de données, nous cliquons sur le bouton « Exécuter ».

Pendant l'exécution du traitement du modèle, l'écran « Etat d'avancement du traitement » (Figure 3-38) s'affiche.

A la fin du traitement, nous cliquons sur « Fermer », et, à l'écran « Traiter Modèle d'exploration de données », nous cliquons également sur « Fermer ».

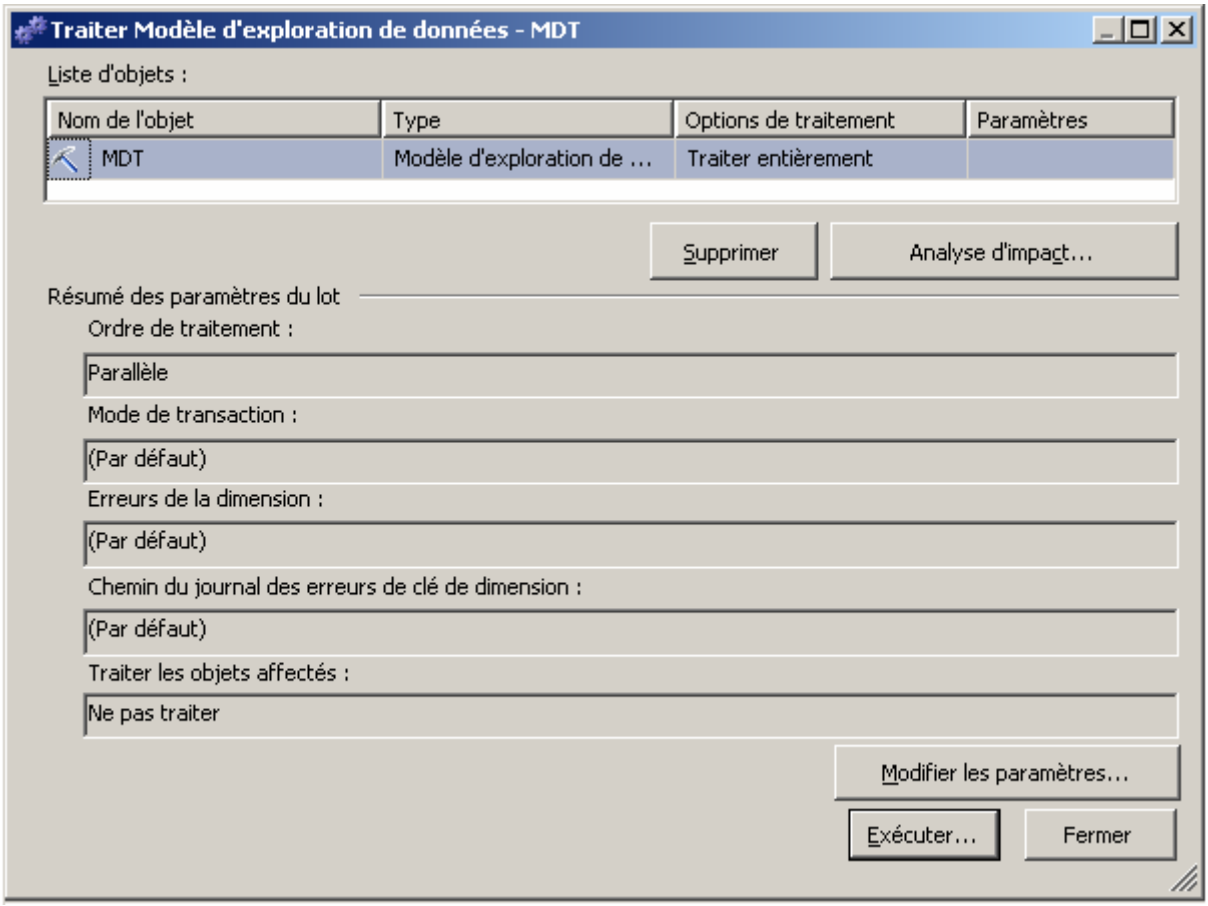


Figure 3-37 : Fenêtre « Publipostage.dmm » - Onglet « Visionneuse d'exploration de données » - Traiter Modèle d'exploration de données

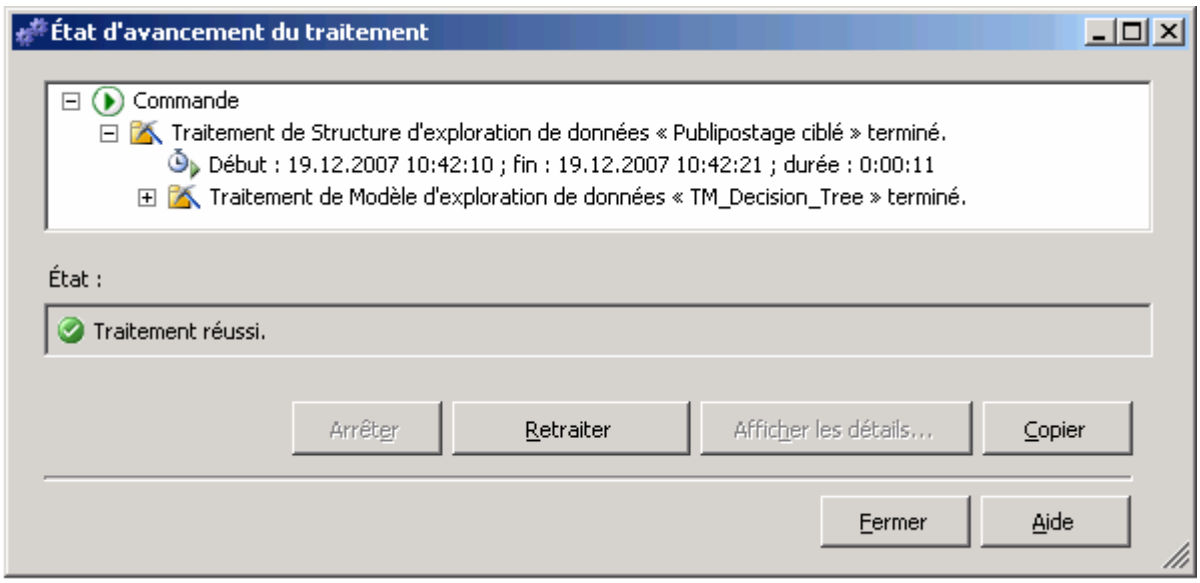


Figure 3-38 : Fenêtre « Publipostage.dmm » - Onglet « Visionneuse d'exploration de données » - Etat d'avancement du traitement

Nous voici enfin sur l'écran « Visionneuse de modèle d'exploration de données » (Figure 3-39).

Nous allons nous attarder sur cet écran car il donne beaucoup d'informations sur le modèle d'exploration de données.

Nous pouvons y observer deux onglets :

- Arborescence de décision ;
- Réseau de dépendances.

3.4.3.1 Arborescence de décision

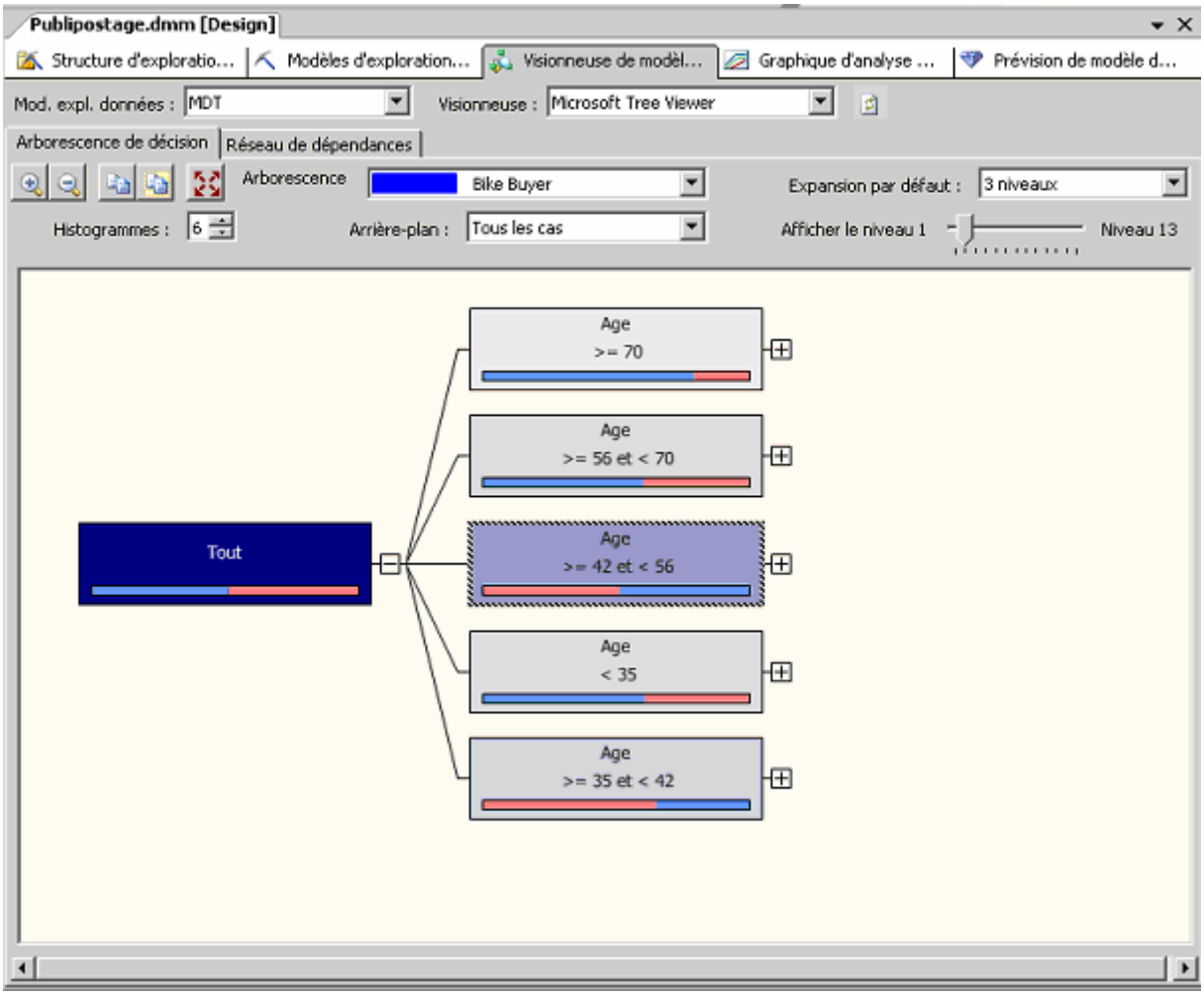


Figure 3-39 : Fenêtre « Publipostage.dmm » - Onglet « Visionneuse d'exploration de données » - Arborescence de décision

L'arborescence d'un arbre de décision débute par les variables regroupant le plus de cas jusqu'à atteindre les niveaux feuilles qui peuvent contenir, par défaut, uniquement 10 cas.
La couleur de remplissage, représente la quantité de cas contenue dans le nœud, et ses descendants, par rapport aux autres nœuds. Plus un nœud est foncé, plus il contient de cas.
Dans notre exemple, le nœud qui contient le plus de personnes est le nœud « Age >= 42 et<56 ».
Nb. Le nœud racine sera toujours le bleu le plus foncé.

A l'intérieur de chaque nœud, nous distinguons une barre de remplissage rouge et bleu. Le coté rouge représente le nombre de cas (en pourcent) de personnes ayant acheté un vélo (« BikeBuyer »=1). Le coté bleu représente donc le nombre de personnes n'ayant pas acheté un vélo (« BikeBuyer »=0).

Le fait de cliquer sur un nœud affiche la « Légende d'exploration de donnée » (Figure 3-40) qui résume le nombre « BikeBuyer » total du nœud, le nombre « BikeBuyer »=1 et le nombre de « BikeBuyer »=0 ainsi que la probabilité pour l'acheteur de faire partie des « BikeBuyer »=1 ou « BikeBuyer »=0.

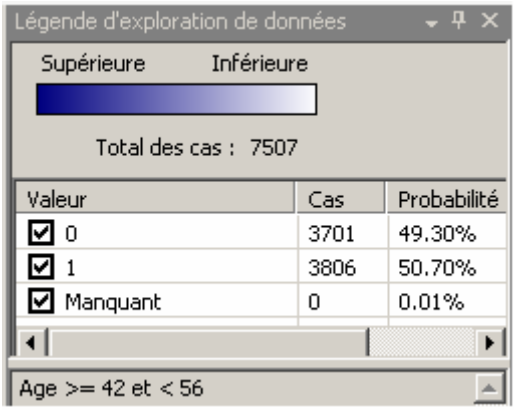


Figure 3-40 : Fenêtre « Publipostage.dmm » - Onglet « Visionneuse d'exploration de données » - Légende d'exploration de données

Au dessus de l'arbre de décision, nous pouvons apercevoir différentes listes box ainsi qu'une barre graduée de 1 à 13. Nous allons y apporter quelques précisions :

Figure 3-41 : Fenêtre « Publipostage.dmm » - Onglet « Visionneuse d'exploration de données » - Liste déroulante « Arborescence »

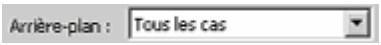


Figure 3-42 : Fenêtre « Publipostage.dmm » - « Visionneuse d'exploration de données » - Liste déroulante « Arrière-plan »

Figure 3-43 : Fenêtre « Publipostage.dmm » - Onglet « Visionneuse d'exploration de données » - Liste déroulante « Expansion par défaut »

Figure 3-44 : Fenêtre « Publipostage.dmm » - Onglet « Visionneuse d'exploration de données » - Afficher le niveau

La liste déroulante « Arborescence » (Figure 3-41) permet, si nous avons définit plusieurs attributs à prédire, de choisir quel est l'arbre de décision que nous désirons afficher, car l'algorithme crée un arbre par valeur à estimer.

La liste déroulante « Arrière-plan » (Figure 3-42) permet d'influencer la couleur d'arrière plan des nœuds : les nœuds affichant des couleurs plus foncées contiennent plus de cas répondant au critère sélectionné. Par exemple, si nous choisissons la valeur 1, le nœud « Number Cars Owned = 0 », suivant le nœud « Age < 34 » possède une couleur très sombre cas nous y trouvons 307 cas dont 267 (86.95%) de cas où « Bike Buyer = 1 »

La liste déroulante « Expansion par défaut » (Figure 3-43) permet de développer par défaut 3, 4, 5 ou tous les nœuds de l'arbre de décision.

La barre graduée « Afficher le niveau » (Figure 3-44) permet de développer les nœuds jusqu'au niveau sélectionné. Dans cet exemple, nous pouvons développer jusqu'à 13 niveaux car l'arbre de décision va jusqu'à un maximum de 13 niveaux.

3.4.3.2 Réseau de dépendance

Cet écran (Figure 3-45) permet de visualiser les liens entre les variables d'entrée et la variable à estimer : plus le lien est fort, plus la variable influence le résultat.

Pour déterminer quelle est la dépendance entre les différents attributs, une réglette à gauche de l'écran permet d'afficher progressivement, en partant depuis le bas et en remontant, les variables qui influencent de moins en moins l'estimation.

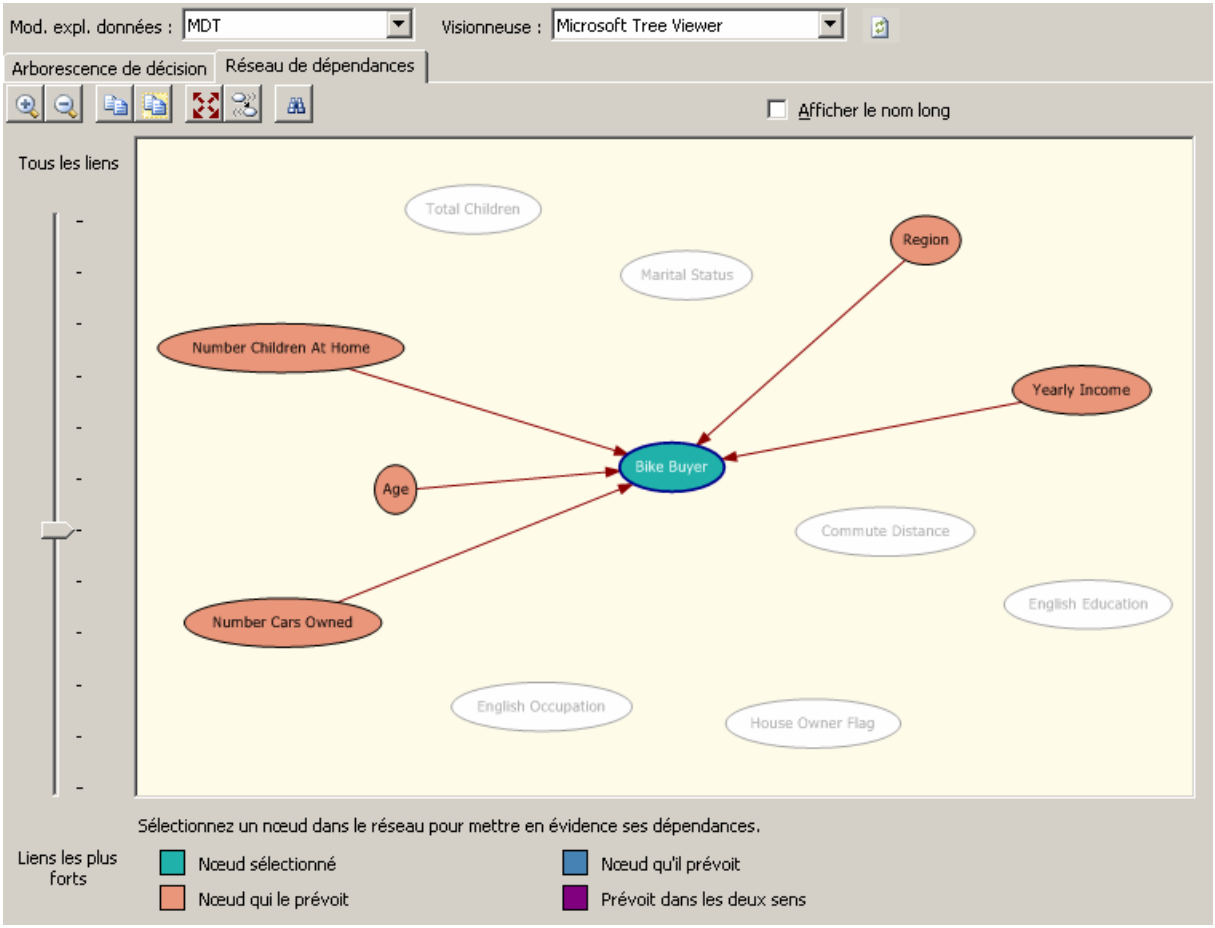


Figure 3-4546 : Fenêtre « Publipostage.dmm » - Onglet « Visionneuse d'exploration de données » - « Réseau de dépendance »

Dans notre exemple, le lien le plus fort est « Number Cars Owned », suivit de « Number Children At Home » ensuite « Age » et ainsi de suite jusqu'à « House Owner Flag », le lien le plus faible.

3.4.4 Graphique d'analyse de précision

Cet onglet (Figure 3-47) permet d'analyser la précision de l'algorithme à l'aide des cas pour lesquels nous connaissons déjà le résultat escompté. Il est intéressant de l'utiliser lorsque nous avons créé plusieurs modèles d'exploration de données afin de les comparer entre eux.

Nous nous attardons sur cet écran dans un prochain chapitre lorsque nous aurons ajoutons des modèles d'exploration de données supplémentaires.

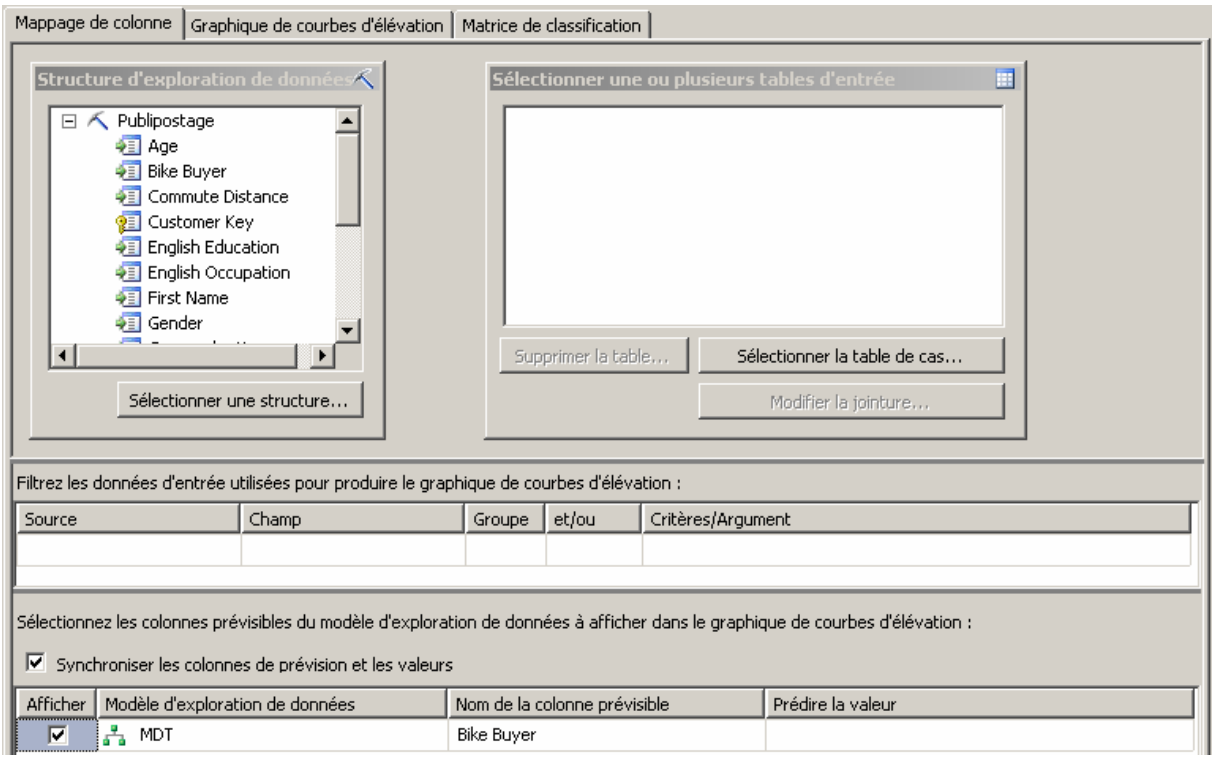


Figure 3-47 : Fenêtre « Publipostage.dmm » - Onglet « Graphique d'analyse de précision »

3.4.5 Prévision de modèles d'exploration de données

Cet onglet (Figure 3-48) permet, lorsque que nous avons choisi un modèle d'exploration de données, d'effectuer des prévisions sur des données « réelles ».

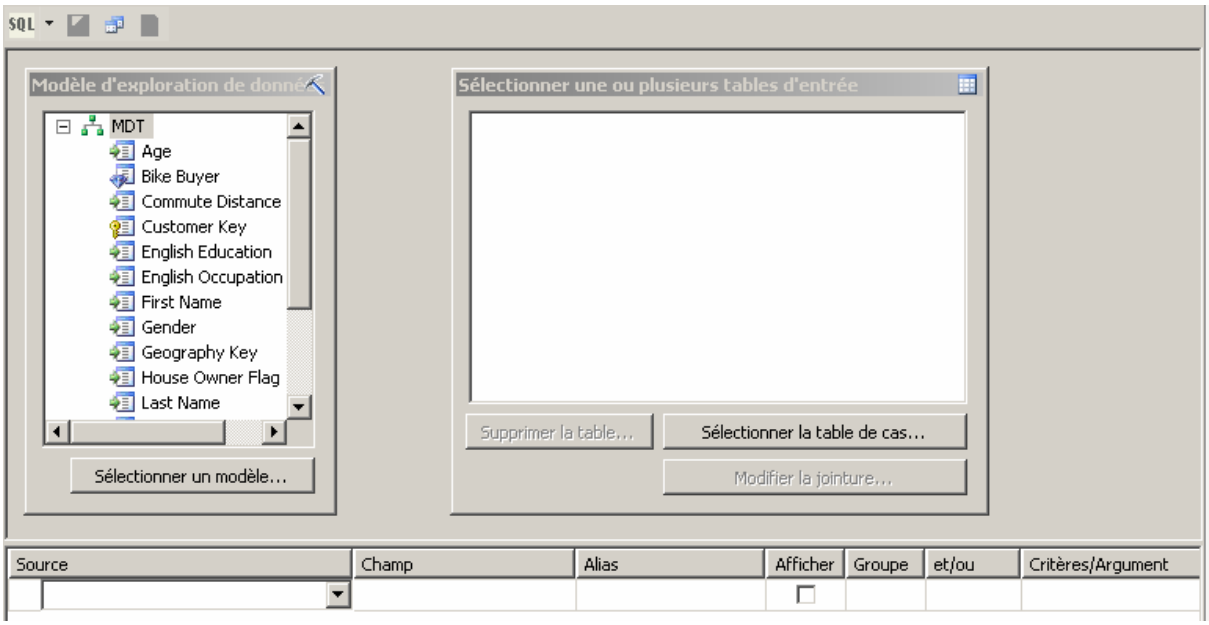


Figure 3-48 : Fenêtre « Publipostage.dmm » - Onglet « Prévision de modèles d'exploration de données »

Cet écran nous permet de générer des requêtes DMX (Data Mining Extensions) à l'aide d'une interface graphique.

La première étape consiste à sélectionner une source de données contenant les cas sur lesquels nous effectuons la prédiction. Nous appelons cette table : Table de cas.
Lors de l'étape « Créer une Vues des sources de données », nous avons sélectionné la table « dbo.ProspectiveBuyer », table qui contient des données sur des acheteurs potentiels de vélo. Nous pourrions aussi créer une autre « Source de donnée » ainsi qu'une autre « Vue de sources de données » qui fait référence à une seconde base de données afin de sélectionner d'autres données provenant des systèmes productifs, par exemple.

Pour sélectionner la « Table de cas », nous cliquons sur le bouton « Sélectionner la table de cas… ». L'action de ce bouton affiche la fenêtre « Sélectionner une table » (Figure 3-49). A partir de cette fenêtre, nous pouvons choisir la « Source de données » contenant la table/vue de laquelle nous extrayons les données que nous désirons prédire.
Etant donné que notre table de cas se trouve dans la source de données « Adventure Works DW », nous sélectionnons la dite source de données et nous choisissons, dans le champ « Nom de la table/vue », la table « ProspectiveBuyer (dbo) ». Nous fermons cette fenêtre en cliquant sur le bouton « OK ».

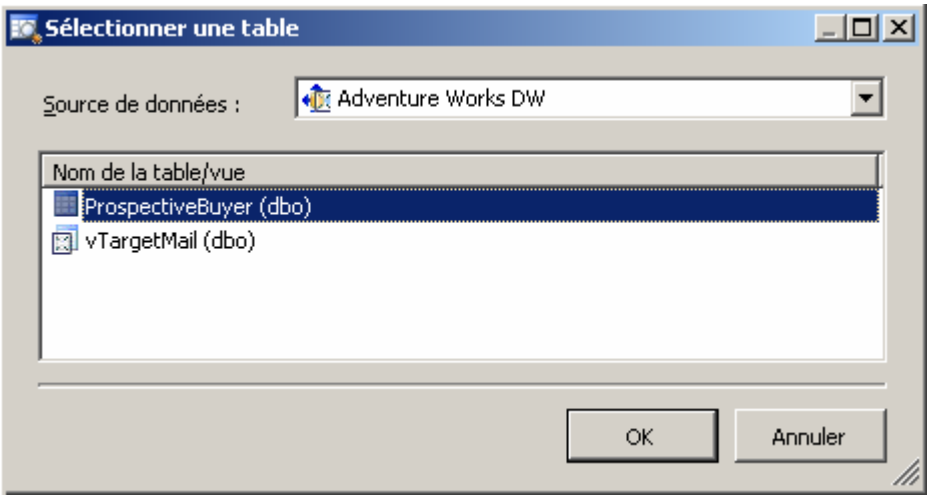


Figure 3-49 : Fenêtre « Publipostage.dmm » - Onglet « Prévision de modèles d'exploration de données » - Sélectionner une table

Par défaut, VS2005 – BI lie le modèle d'exploration de données et la table de cas avec les noms de champs correspondant d'une table à l'autre.
Dans notre exemple, la table des cas ayant des noms de colonne correspondant au modèle d'exploration de donnée (« Age », « FirstName », « Gender », « HouseOwnerFlag », « LastName », « MaritalStatus », « NumberCasOwned », « NumberChildrenAtHome », « TotalChildren » et « YearlyIncome ») ont été automatiquement liés (Figure 3-50).

Si nous nous référons au réseau de dépendance de la Figure 3-45, nous nous constatons que certains champs se sont liés alors qu'ils ne sont pas influant et que d'autres manquent.
Nous devons donc « délier » les champs « FirstName » et « LastName » en cliquant sur leur lien respectif et en appuyant sur la touche « Delete » du clavier.
Ensuite, nous devons « Lier » les champs « English Education » avec « Education » et « English Occupation » avec « Occupation » en faisant un glisser/déplacer depuis le champ sélectionné du modèle d'exploration de données jusqu'au champ correspondant de la table des cas.
Le champ « Commute Distance », mentionné dans le réseau de dépendance, n'étant pas disponible dans notre table de cas, nous ne pouvons pas le lier.

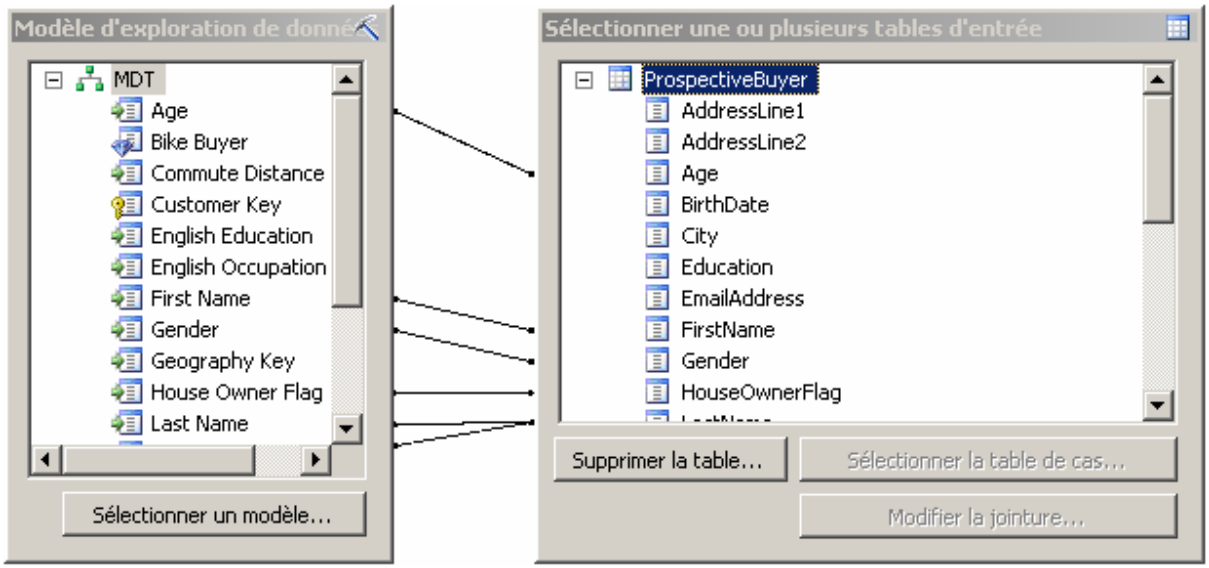


Figure 3-50 : Fenêtre « Publipostage.dmm » - Onglet « Prévision de modèles d'exploration de données » - Edition des jointures

La seconde étape consiste à spécifier quelle est la variable à prédire (dans l'éventualité où il y en a plusieurs), quelle est la clé de la table des cas ainsi que la

fonction de prévision qui devra être appliqué. Ces spécifications sont à saisir dans la deuxième partie de la Figure 3-48.

Pour spécifier la variable à prédire, dans le champ « Source » nous spécifions quel est le modèle de prédiction choisis (ici « MDT ») et dans la colonne « Champ » la variable à prédire (ici « Bike Buyer »).
Pour spécifier la clé de la table de cas, dans le champ « Source » nous choisissons dans la liste déroulant « table ProspectiveBuyer » et, dans la liste déroulant « Champ », la variable « ProspectAlternateKey ».
Pour finir, nous spécifions « Fonction de prévision » à partir de la liste déroulante de la colonne « Source » et, dans la colonne « Champ », « PrédictProbabilité ». Il est nécessaire, lorsque nous spécifions une « Fonction de prévision », d'indiquer quelle est la variable à prendre en compte dans la fonction (ici « [MDT].[Bike Buyer] ».

Le résultat de ces choix est représenté dans la Figure 3-51.

Source	Champ	Alias	Afficher	Groupe	et/ou	Critères/Argument
MDT	Bike Buyer		☑			
ProspectiveBuyer	ProspectAlternateKey		☑			
Fonction de prévision	PredictProbability		☑			[MDT].[Bike Buyer]
			☐			

Figure 3-51 : Fenêtre « Publipostage.dmm » - Onglet « Prévision de modèles d'exploration de données » - Paramétrage de la requête de prédiction

En cliquant sur la liste déroulante au sommet à gauche (Figure 3-52), nous pouvons sélectionner « SQL Requête » afin de consulter la requête SQL (DMX) générée par VS2005 –BI (Script 3-1).

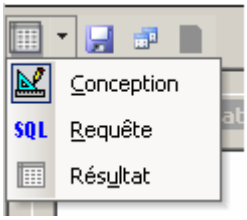


Figure 3-52 : Fenêtre « Publipostage.dmm » - Onglet « Prévision de modèles d'exploration de données » - Liste déroulante

```
SELECT
  [MDT].[Bike Buyer],
  t.[ProspectAlternateKey],
  PredictProbability([MDT].[Bike Buyer])
From
  [MDT]
PREDICTION JOIN
  OPENQUERY([Adventure Works DW],
    'SELECT
      [ProspectAlternateKey],
      [MaritalStatus],
      [Gender],
      [YearlyIncome],
      [TotalChildren],
      [NumberChildrenAtHome],
      [HouseOwnerFlag],
      [NumberCarsOwned],
      [Age],
      [Education],
      [Occupation]
    FROM
      [dbo].[ProspectiveBuyer]
```

```
    ') AS t
ON
  [MDT].[Marital Status] = t.[MaritalStatus] AND
  [MDT].[Gender] = t.[Gender] AND
  [MDT].[Yearly Income] = t.[YearlyIncome] AND
  [MDT].[Total Children] = t.[TotalChildren] AND
  [MDT].[Number Children At Home] = t.[NumberChildrenAtHome] AND
  [MDT].[House Owner Flag] = t.[HouseOwnerFlag] AND
  [MDT].[Number Cars Owned] = t.[NumberCarsOwned] AND
  [MDT].[Age] = t.[Age] AND
  [MDT].[English Education] = t.[Education] AND
  [MDT].[English Occupation] = t.[Occupation]
```

Script 3-1

Dans la liste déroulante de la Figure 3-52, nous choisissons cette fois « Résultat » () afin que SSAS traite les informations de la table de cas et qu'il nous retourne les résultats sous forme de liste (Figure 3-53).

En faisant un clic droit sur cette liste, nous pouvons choisir l'option du menu contextuel « Copier tout » (Figure 3-54) afin de traiter, par exemple, ces résultats dans MS Excel.

Bike Buyer	ProspectAlterna...	Expression
1	827	0.63383639888...
0	833	0.52113422833...
0	844	0.73569722117...
0	832	0.91217704030...
0	53313373327	0.54435760318...
0	54107006788	0.50529701670...
1	53315894603	0.76055638350...
1	54037360548	0.67956940485...
1	15732240080	0.67889717488...
1	53469896316	0.70920520106...
0	15737550900	0.74929159094...
0	21596444800	0.65150381409...
1	75533120036	0.67956940485...
0	21534748700	0.55026367376...
0	21585115194	0.65150381409...
0	21662647128	0.55026367376...
1	68392822146	0.63383639888...
1	69732639193	0.63383639888...
0	1242	0.68791317971...
0	3003	0.50529701670...
0	3014	0.65150381409...
1	2997	0.67889717488...
0	3001	0.65150381409...
0	3002	0.68791317971...
1	3004	0.63383639888...

Figure 3-53 : Fenêtre « Publipostage.dmm » - Onglet « Prévision de modèles d'exploration de données » - Résultat de l'estimation

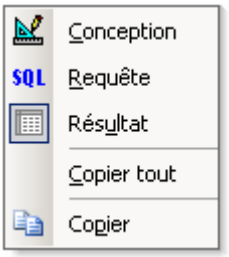


Figure 3-54 : Fenêtre « Publipostage.dmm » - Onglet « Prévision de modèles d'exploration de données » - Menu contextuel

Il existe aussi la possibilité de les enregistrer directement dans une table d'une base de données en cliquant sur l'icône de la disquette ( ) à droite de la liste déroulante de la Figure 3-52.
Ce clic affiche la fenêtre « Enregistrer le résultat de la requête d'exploration de données » (Figure 3-55).

A partir de cette fenêtre, il est nécessaire de choisir dans quelle « Source de données » (ici « Adventure Works DW ») nous désirons exporter les données, donner le nom d'une table cible (si la table n'existe pas dans la base de données, VS2005 – BI créera la table) et, si par exemple nous désirons traiter les prédictions dans un autre modèle d'exploration de donnée, ajouter la table contenant les données dans une « Vue de source de données » existante via la liste déroulante « Ajouter à la vue de source de données ».
La case à cocher « Remplacer en cas d'existence » fait un « Drop Table » de la table cible si celle-ci existe déjà dans la base de données.

L'enregistrement s'exécute au moment du clic sur le bouton « Enregistrer ».

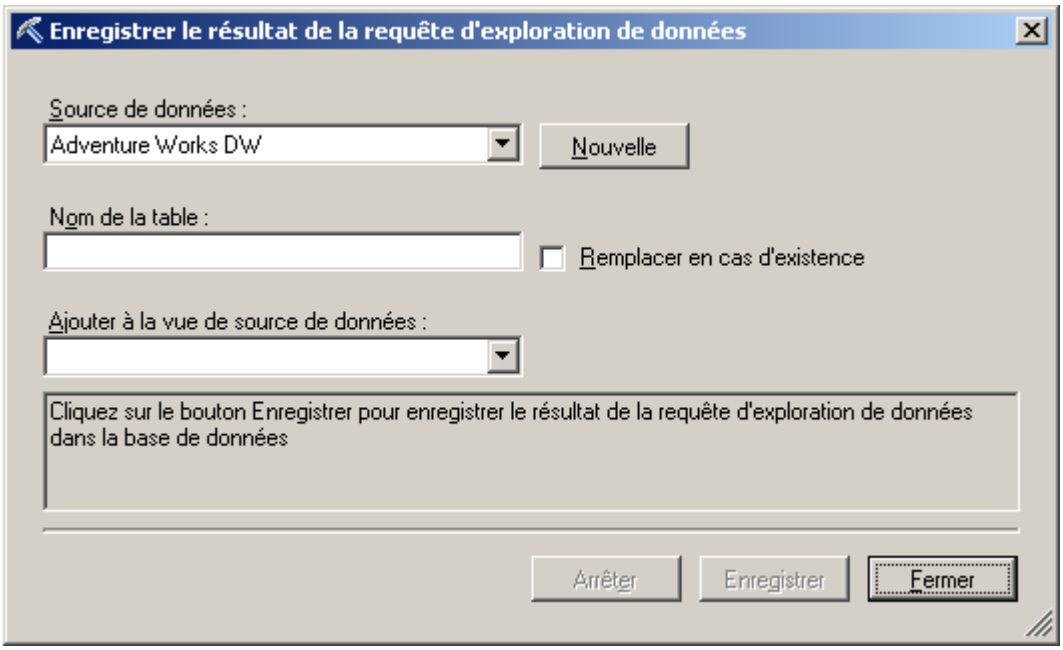


Figure 3-55 : Fenêtre « Publipostage.dmm » - Onglet « Prévision de modèles d'exploration de données » - Enregistrer le résultat de la requête d'exploration de données

Dans la Figure 3-52, nous apercevons qu'il reste encore une icône à expliquer ( ) : la requête « Singleton ».
Une requête Singleton permet de fournir au modèle d'exploration de données les informations relatives à un cas dans le cadre, par exemple, d'un devis pour une facture.
Nous cliquons sur cette dernière icône et nous pouvons apercevoir que la table de cas de la Figure 3-48 change pour devenir la table « Entrée de requête singleton » (Figure 3-56).

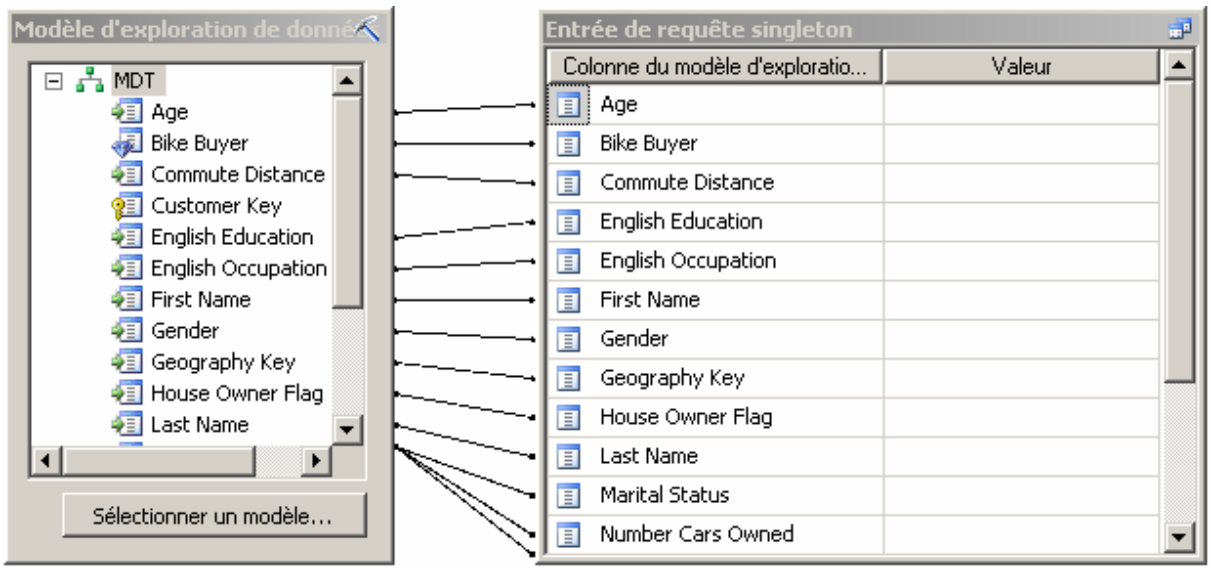


Figure 3-56 : Fenêtre « Publipostage.dmm » - Onglet « Prévision de modèles d'exploration de données » - Requête Singleton

Nous allons tester cette méthode en choisissant, depuis les listes déroulantes disponibles dans la colonne « Valeur », les variables nécessaires à cette simulation, selon le Tableau 3-2.

Ces listes déroulantes sont construites à partir des valeurs des données de tests qui ont permis l'apprentissage du modèle d'exploration de données.

	Valeur
Age	39-48
Bike Buyer	
Commute Distance	5-10 Miles
English Education	Bachelors
English Occupation	Clerical
First Name	
Gender	F
Geography Key	
House Owner Flag	1
Last Name	
Marital Status	S
Number Cas Owned	1
Number Children At Home	1
Region	
Total Children	1
Yearly Income	30000

Tableau 3-2 : Valeur des variables pour la requête Singleton

Nous ignorons volontairement les champs « First Name », « Last Name », « Bike Buyer » et « Region » car ces champs ne sont pas utiles dans le fonctionnement de l'algorithme.

Ensuite, il est nécessaire, comme dans le cas d'une table de cas, de spécifier quel est le modèle d'exploration de donnée ainsi que la variable sur laquelle nous désirons effectuer cette simulation. Nous reproduisons la Figure 3-51 en ignorant la liste contenant la ligne spécifiant la clé de la table de cas qui n'existe pas ici.

Nous choisissons, dans le menu déroulant de la Figure 3-54, l'option « Résultat » ( ) afin que SSAS applique le modèle d'exploration de données et qu'il nous retourne le résultat (Figure 3-57).

Il en ressort que cette acheteuse potentielle est une acheteuse de vélo pour taux de probabilité de 76.47 %.

Bike Buyer	Expression
1	0.76469398033236546

Figure 3-57 : Fenêtre « Publipostage.dmm » - Onglet « Prévision de modèles d'exploration de données » - Résultat requête Singleton

3.5 Navigation dans une Structures d'exploration de données – MNB

Dans le chapitre précédent (Navigation dans une Structures d'exploration de données – MDT), nous avons étudié le modèle d'exploration de données « Microsoft Decision Tree ». Dans ce chapitre, nous nous intéressons à la mise en place et à la navigation d'un modèle d'exploration de données basé sur l'algorithme « Microsoft Naive Bayes »

Nous plaçons ce nouveau modèle d'exploration de données dans la même structure d'exploration de données que celle spécifiée dans le chapitre précédent, c'est à dire : « Publipostage.dmm ». Pour ajouter ce modèle, nous faisons un clic droit sur la structure d'exploration et nous choisissons, dans le menu contextuel, l'option « Ouvrir » (Figure 3-58).

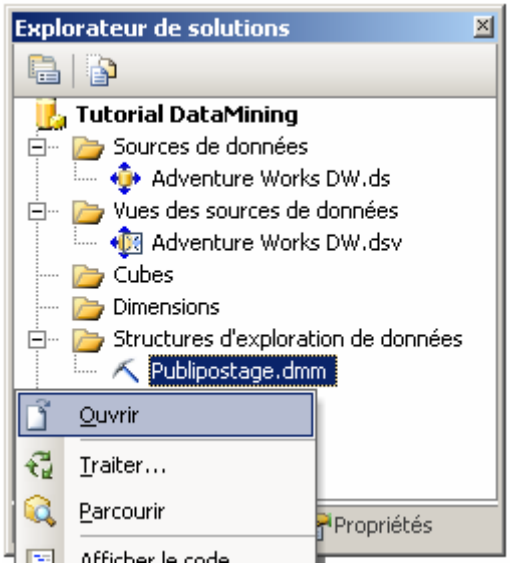


Figure 3-58 : Fenêtre « Explorateur de solution » - Ouvrir… MNB

Une fois la structure « Publipostage.dmm » ouverte, nous choisissons l'onglet « Modèles d'exploration de données » (Figure 3-59) et nous cliquons sur l'icône ( ) qui signifie « Créer un modèle d'exploration connexe ».

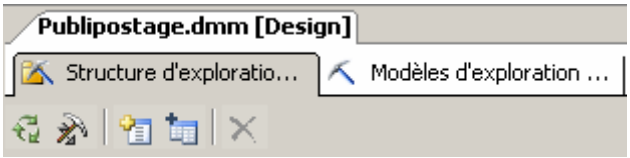


Figure 3-59 : Fenêtre « Publipostage.dmm » - Onglet « Structure d'exploration de données »

L'écran « Nouveau modèle d'exploration de données » (Figure 3-60) s'affiche. Dans celui-ci, nous saisissons, comme « Nom du modèle » MNB et nous choisissons dans la liste déroulante « Nom d'algorithme » : « Algorithme MNB (Microsoft Naive Bayes) » et nous cliquons sur le bouton « OK ».

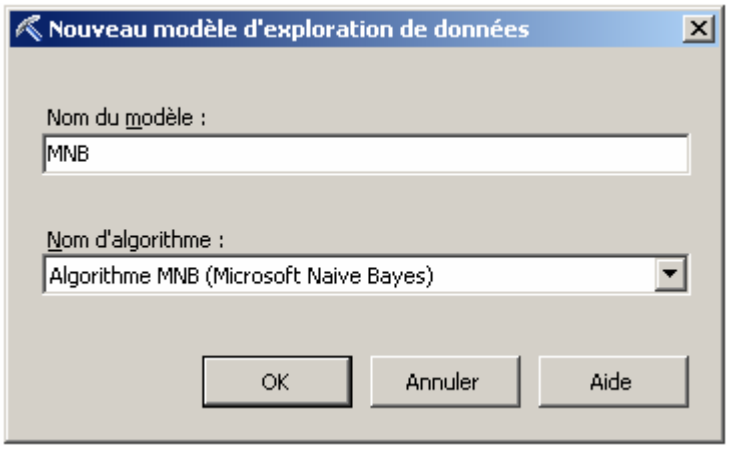


Figure 3-60 : Fenêtre « Publipostage.dmm » - Onglet « Structure d'exploration de données » - Nouveau modèle d'exploration de données

Etant donné que l'« Algorithme MNB (Microsoft Naive Bayes) » ne prend pas en charge les champs ayant un type de contenu « continuous (continue) », SSAS nous rappelle ce détail en affichant le message de la Figure 3-61, difficilement compréhensible, et nous informe que la variable « Yearly Income » est ignorée. Nous acceptons d'ignorer cette colonne en cliquant sur « Oui ».

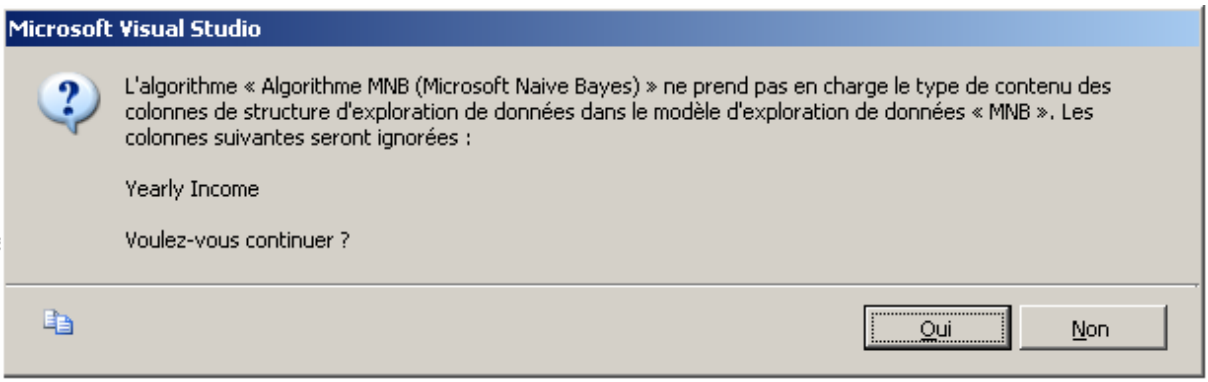


Figure 3-61 : Fenêtre « Publipostage.dmm » - Onglet « Structure d'exploration de données » - Boîte d'information « Type de contenu »

Comme affiché sur la Figure 3-62, notre structure d'exploration de données « Publipostage.dmm » s'est enrichie du nouveau modèle d'exploration de données.
Nous pouvons également observer sur cet écran que la variable « Yearly Income », comme indiqué précédemment, est ignorée.

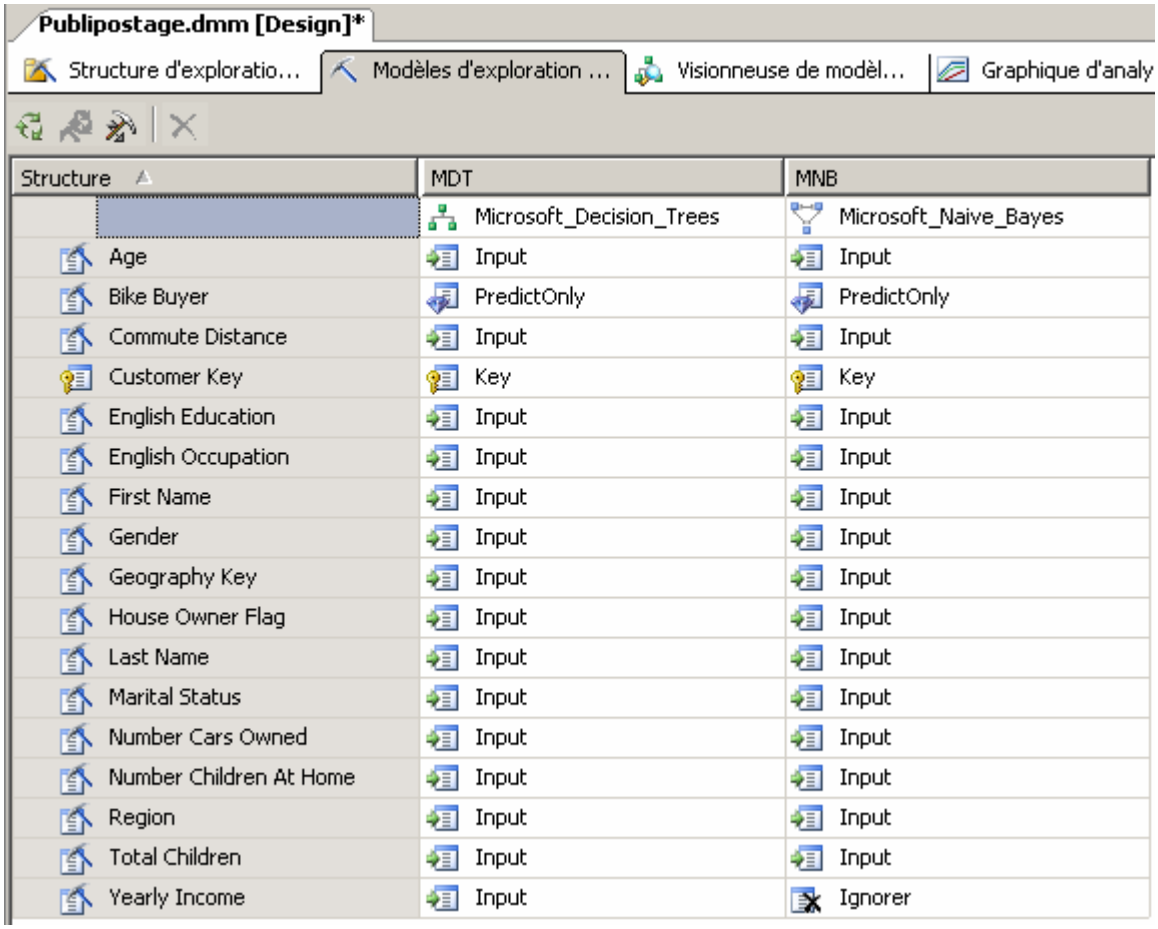


Figure 3-62 : Fenêtre « Publipostage.dmm » - Onglet « Structure d'exploration de données » - Deux modèles d'exploration de données

Avant de continuer, il est nécessaire de traiter ce nouveau modèle. Nous le faisons en cliquant sur l'icône « Traiter la structure d'exploration de donnée et l'ensemble de ses modèles associés » (). Nous ferons l'impasse sur la Figure 3-37 ainsi que la Figure 3-38 et leurs actions associées.

3.5.1 Visionneuse de modèle d'exploration de données

Une fois l'apprentissage effectué, nous cliquons sur l'onglet « Visionneuse de modèle d'exploration de données » (Figure 3-63).
Par rapport à la visionneuse de l'algorithme « Microsft Decision Tree », nous distinguons quatre nouveaux sous onglets :

- Réseau de dépendance ;
- Profil d'attribut ;
- Caractéristique d'attribut ;
- Discrimination d'attribut.

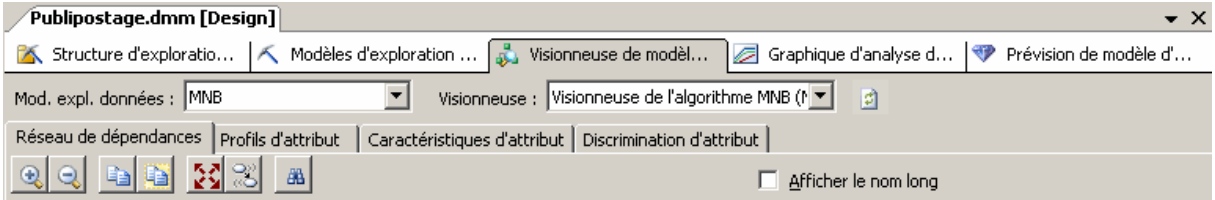


Figure 3-63 : Fenêtre « Publipostage.dmm » - Onglet « Visionneuse d'exploration de données » - Algorithme MNB

3.5.1.1 Réseau de dépendance

Le réseau de dépendance du modèle d'exploration de donnée « Microsoft Naive Bayes » fonctionne exactement comme le réseau de dépendance de l'algorithme « Microsoft

Decision Tree » décrit au chapitre « 3.4.3.2 Réseau de dépendance ». Nous ne nous attardons donc pas à nouveau sur ce chapitre.

3.5.1.2 Profils d'attribut

Cet onglet (Figure 3-64) présente la liste des attributs (1 ligne par attribut dans la colonne « Attributs ») ainsi que les différents états possibles de la variable à prédire (ici « Bike Buyer »).

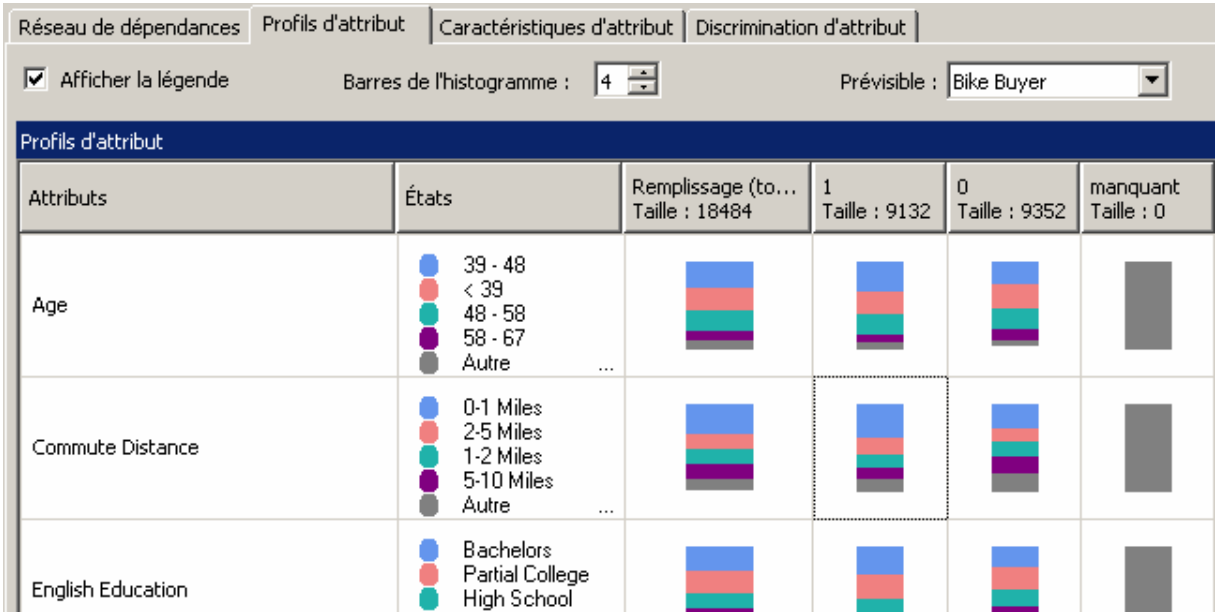


Figure 3-64 : Fenêtre « Publipostage.dmm » - Onglet « Visionneuse d'exploration de données » Profils d'attribut

A l'intersection d'une ligne d'attribut et d'une colonne, nous pouvons observer la distribution des cas et, lorsque nous cliquons sur l'un de ces croisements, SSAS nous affiche la « Légende d'exploration de données » (Figure 3-65) qui nous renseigne sur la probabilité, pour un état de l'attribut, de prendre la valeur spécifiée dans l'entête de colonne.
Par exemple, dans la Figure 3-64 nous avons cliqué à l'intersection de la ligne « Commute Distance » et de la colonne 1 (« Bike Buyer = 1 »). Le résultat de ce croisement est résumé dans la Figure 3-65 et nous pouvons observer que si l'acheteur potentiel habite à moins d'un mile de son lieu de travail, la probabilité qu'il achète un vélo est de 38.7 %

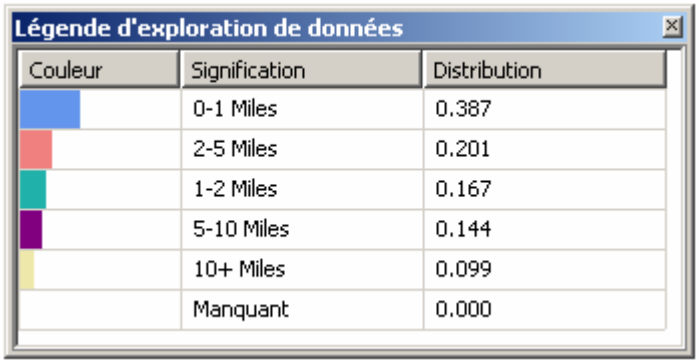


Figure 3-65 : Fenêtre « Publipostage.dmm » - Onglet « Visionneuse d'exploration de données » - Profils d'attribut - Légende d'exploration de données

Les options disponibles durant l'affichage de cet onglet sont :

- Afficher la légende : cette option permet d'afficher ou de masquer la colonne « Etats »
- Barres de l'histogramme : 4 : cette option permet d'ajouter ou de supprimer dans les histogrammes des croisements un état de l'attribut

3.5.1.3 Caractéristique d'attribut

Cet onglet (Figure 3-66) permet de classer par ordre décroissant les états des attributs selon la valeur attendue pour la variable sélectionnée (ici « Bike Buyer = 1 ».

Nous pouvons en déduire que les acheteurs de vélos ont une probabilité de ne pas avoir d'enfant à la maison (62.73 %), d'habiter à moins d'un mile du travail (38.74 %), etc.
Nous obtenons le pourcentage simplement en pointant la souris sur le croisement de la ligne d'un attribut et la colonne « Probabilité ».

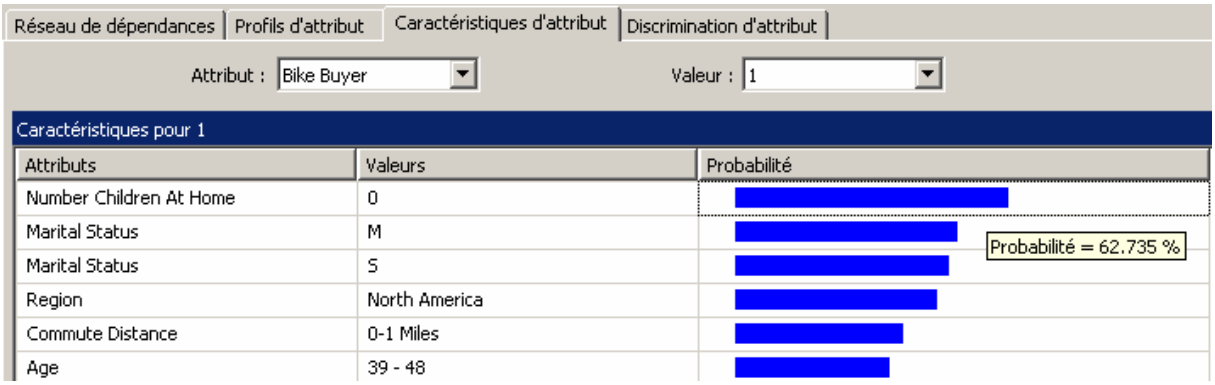


Figure 3-66 : Fenêtre « Publipostage.dmm » - Onglet « Visionneuse d'exploration de données » - Caractéristique d'attribut

3.5.1.4 Discrimination d'attribut

Cet onglet (Figure 3-67) permet de comparer la valeur d'un attribut par rapport à une autre valeur de cet attribut en fonction de la variable sélectionnée.

Par exemple, nous choisissons comme « Valeur 1 : 1 » (« Bike Buyer »=1) et comme « Valeur 2 : 0 » (« Bike Buyer »=0).
Nous pouvons apercevoir que lorsque nous comparons l'attribut « Number Casr Owned » selon qu'il est égal à 0 (pas de voiture) ou à 2, les acheteurs de vélo n'ont pas de voiture alors que ceux qui possèdent 2 voitures n'en achètent pas.

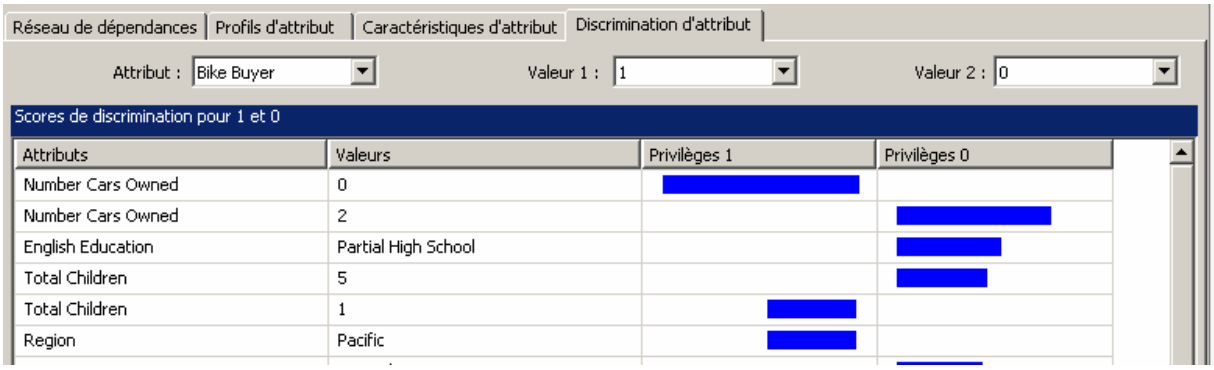
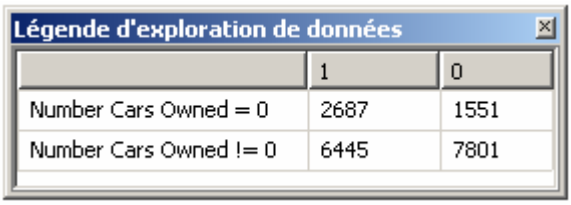


Figure 3-6768 : Fenêtre « Publipostage.dmm » - Onglet « Visionneuse d'exploration de données » - Discrimination d'attribut

Lorsque nous cliquons sur une ligne d'un attribut, la « Légende d'exploration de données » s'affiche (Figure 3-69) et résume les cas contenus dans le nœud.

Par exemple, lorsque nous cliquons sur le nœud « Number Cars Owned », SSAS nous indique qu'il y a 2687 acheteurs de vélos qui ne possèdent pas de voiture et qu'il y a 1551 personnes qui, n'ayant pas acheté de vélos, ne possèdent pas de voiture non plus.



	1	0
Number Cars Owned = 0	2687	1551
Number Cars Owned != 0	6445	7801

Figure 3-69 : Fenêtre « Publipostage.dmm » - Onglet « Visionneuse d'exploration de données » - Discrimination d'attribut - Légende d'exploration de données

3.6 Navigation dans une Structures d'exploration de données – MC

Dans les chapitres précédents (3.4 Navigation dans une Structures d'exploration de données – MDT et 3.5 Navigation dans une Structures d'exploration de données – MNB), nous avons étudiés les modèles d'exploration de données « Microsoft Decision Tree » et « Microsoft Naive Bayes ». Dans ce chapitre, nous nous intéressons à la mise en place et à la navigation d'un modèle d'exploration de données basé sur l'algorithme « Cluster Microsoft »

Nous plaçons ce nouveau modèle d'exploration de données dans la même structure d'exploration de données que celle spécifiée dans les chapitres précédents, c'est à dire : « Publipostage.dmm ». Pour ajouter ce modèle, nous faisons un clic droit sur la structure d'exploration et nous choisissons, dans le menu contextuel, l'option « Ouvrir » (Figure 3-70).

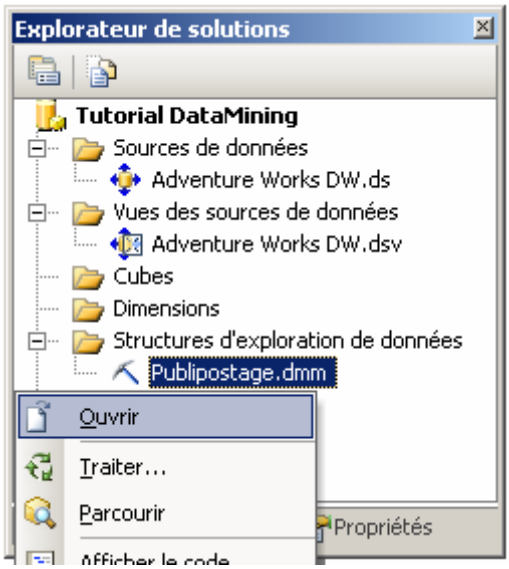


Figure 3-70 : Fenêtre « Explorateur de solutions » - Ouvrir… MC

Une fois la structure « Publipostage.dmm » ouverte, nous choisissons l'onglet « Modèles d'exploration de données » (Figure 3-71) et nous cliquons sur l'icône () qui signifie « Créer un modèle d'exploration connexe ».

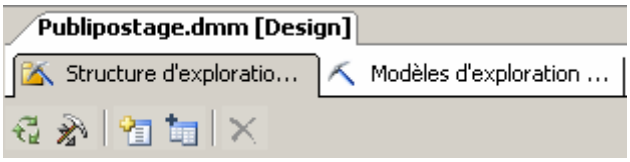


Figure 3-71

L'écran « Nouveau modèle d'exploration de données » (Figure 3-60) s'affiche. Dans celui-ci, nous saisissons, comme « Nom du modèle », MC et nous choisissons dans la liste déroulante « Nom d'algorithme » : « Clusters Microsoft » et nous cliquons sur le bouton « OK ».

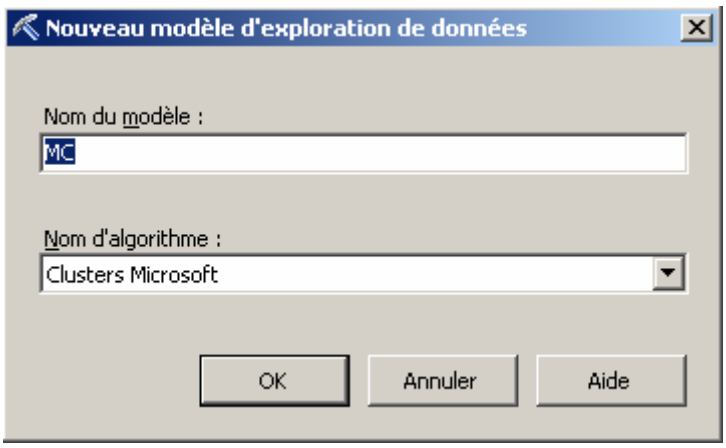


Figure 3-72 : Fenêtre « Publipostage.dmm » - Onglet « Modèles d'exploration de données » - Nouveau modèle d'exploration de données

Comme affiché sur la Figure 3-73, notre structure d'exploration de données « Publipostage.dmm » s'est doté du nouveau modèle d'exploration de données.

Publipostage.dmm [Design]

Structure d'exploration de données | Modèles d'exploration de données | Visionneuse de modèle d'explora... | Graphique d'analyse de pr

Structure	MDT	MNB	MC
	Microsoft_Decision_Trees	Microsoft_Naive_Bayes	Microsoft_Clustering
Age	Input	Input	Input
Bike Buyer	PredictOnly	PredictOnly	PredictOnly
Commute Distance	Input	Input	Input
Customer Key	Key	Key	Key
English Education	Input	Input	Input
English Occupation	Input	Input	Input
First Name	Input	Input	Input
Gender	Input	Input	Input
Geography Key	Input	Input	Input
House Owner Flag	Input	Input	Input
Last Name	Input	Input	Input
Marital Status	Input	Input	Input
Number Cars Owned	Input	Input	Input
Number Children At Home	Input	Input	Input
Region	Input	Input	Input
Total Children	Input	Input	Input
Yearly Income	Input	Ignorer	Input

Figure 3-73 : Fenêtre « Publipostage.dmm » - Onglet « Modèles d'exploration de données » - trois modèles d'exploration de données

Avant de continuer, il est nécessaire de traiter ce nouveau modèle. Nous le faisons en cliquant sur l'icône « Traiter la structure d'exploration de donnée et l'ensemble de ses modèles associés » ( ). Nous faisons l'impasse sur la Figure 3-37 ainsi que la Figure 3-38 et leurs actions associées.

3.6.1 Visionneuse de modèle d'exploration de données

Une fois l'apprentissage effectué, nous cliquons sur l'onglet « Visionneuse de modèle d'exploration de données » (Figure 3-74).
Par rapport aux visionneuses précédentes, nous distinguons quatre nouveaux sous onglets :

- Diagramme de cluster
- Profil de cluster
- Caractéristique du cluster
- Discrimination du cluster

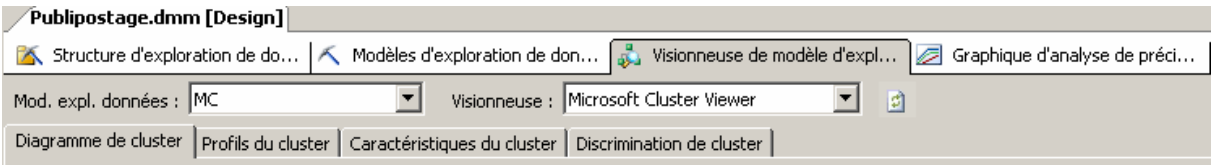


Figure 3-74 : Fenêtre « Publipostage.dmm » - Onglet « Visionneuse de modèles d'exploration de données »

3.6.1.1 Diagramme de cluster

Le diagramme de cluster (Figure 3-75) affiche la dépendance des clusters entre les un et les autres. Plus le lien est important, plus le lien est représenté en gras. Comme dans les réseau de dépendance, une réglette sur la droite permet d'afficher, de bas en haut, les liens les plus fort jusqu'aux liens les plus faibles.
La couleur de remplissage indique, pour la variable et son état choisi, le pourcentage du nombre de cas contenu dans le cluster (ici, le cluster le plus important est le cluster 6).

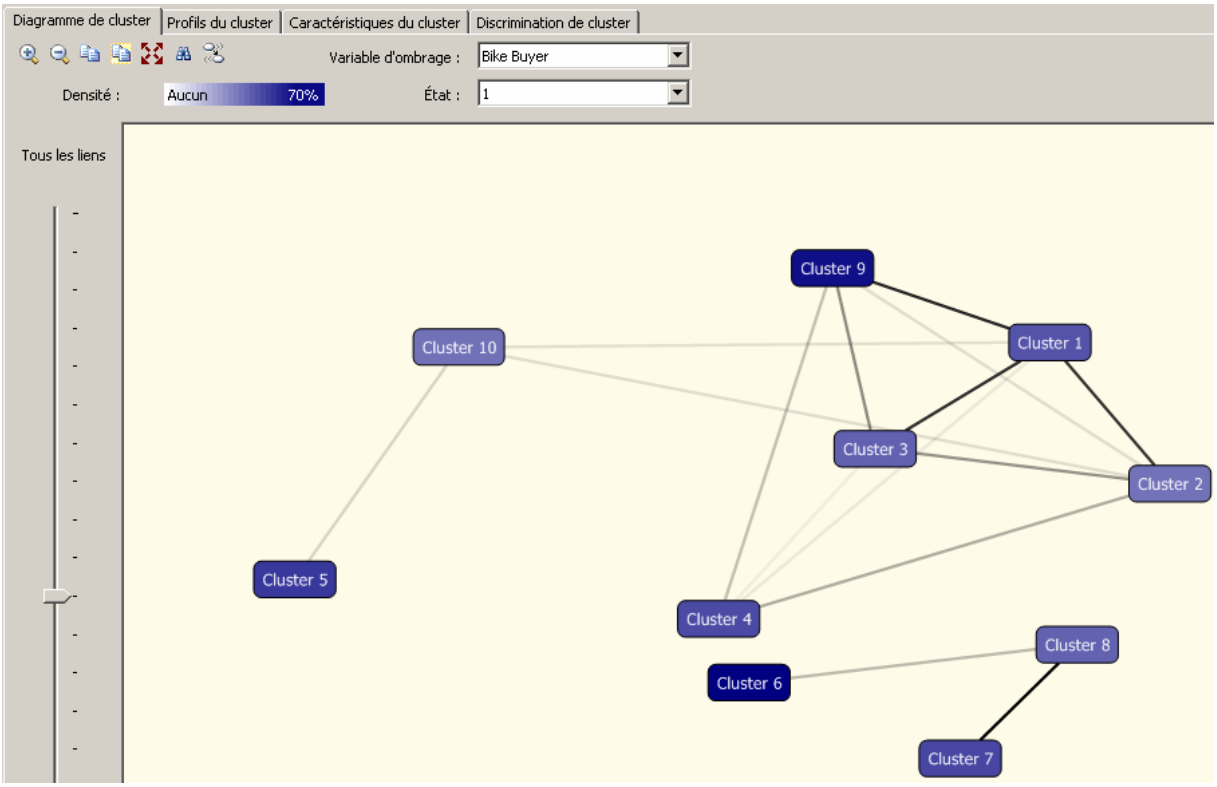


Figure 3-75 : Fenêtre « Publipostage.dmm » - Onglet « Visionneuse de modèles d'exploration de données » Diagramme de cluster

3.6.1.2 Profil du cluster

Comme pour le « Profil d'attribut » du modèle d'exploration « Microsoft Naive Bayes », cet onglet (Figure 3-76) présente la liste des attributs (1 ligne par attribut dans la colonne « Attributs ») ainsi que les différents états possibles contenus dans les clusters disposés en colonne.

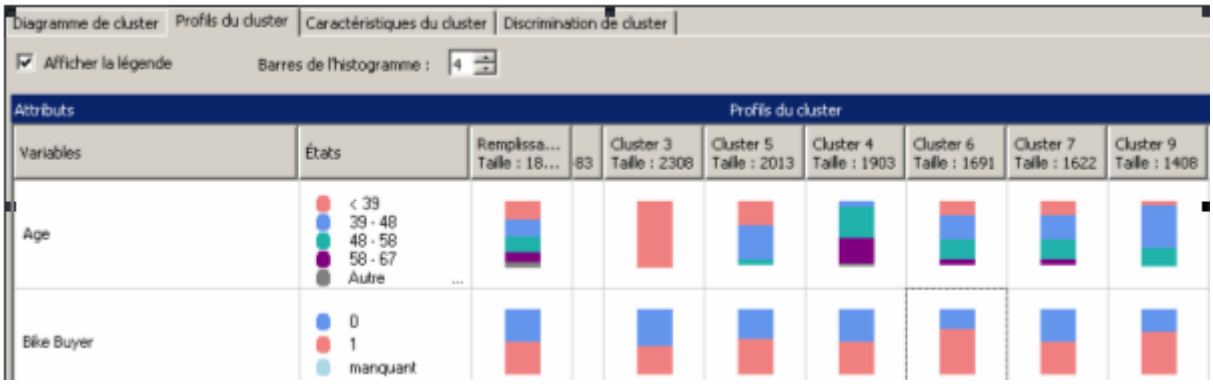


Figure 3-76 : Fenêtre « Publipostage.dmm » - Onglet « Visionneuse de modèles d'exploration de données » Profil de cluster

A l'intersection d'une ligne d'attribut et d'une colonne, nous pouvons observer la distribution des cas et, lorsque nous cliquons sur l'un de ces croisements, SSAS nous affiche la « Légende d'exploration de données » (Figure 3-77) qui nous renseigne sur la probabilité, pour un état de l'attribut, de prendre la valeur spécifiée dans l'entête de colonne.
Par exemple, dans la Figure 3-76 nous avons cliqué à l'intersection de la ligne « Bike Buyer » et de la colonne du cluster 6. Le résultat de ce croisement est résumé dans la Figure 3-77 et nous pouvons observer dans le cluster 6, que les cas ayant acheté un vélo représente 69.9 % des cas du cluster.

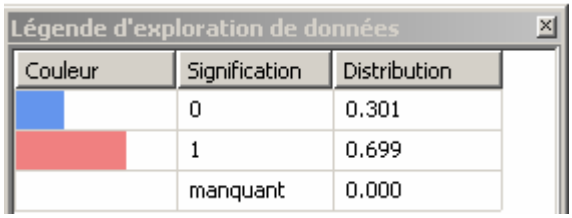
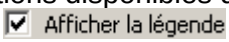


Figure 3-77 : Fenêtre « Publipostage.dmm » - Onglet « Visionneuse de modèles d'exploration de données » Profil de cluster – Légende d'exploration de données

Les options disponibles durant l'affichage de cet onglet sont :

- 

: cette option permet d'afficher ou de masquer la colonne « Etats »
- 

: cette option permet d'ajouter ou de supprimer dans les histogrammes des croisements un état de l'attribut

3.6.1.3 Caractéristique du cluster

Comme pour la « Caractéristique d'attribut » du modèle d'exploration « Microsoft Naive Bayes », cet onglet (Figure 3-78) permet de classer par ordre décroissant les états des variables selon le cluster sélectionné

Nous pouvons en déduire que les cas contenus dans ce cluster ont une forte probabilité de ne pas avoir d'enfant à la maison (81.17 %), ni d'habiter à moins d'un mile du travail (87.45 %), etc.
Nous obtenons le pourcentage simplement en pointant la souris sur le croisement de la ligne d'un attribut et de la colonne « Probabilité ».

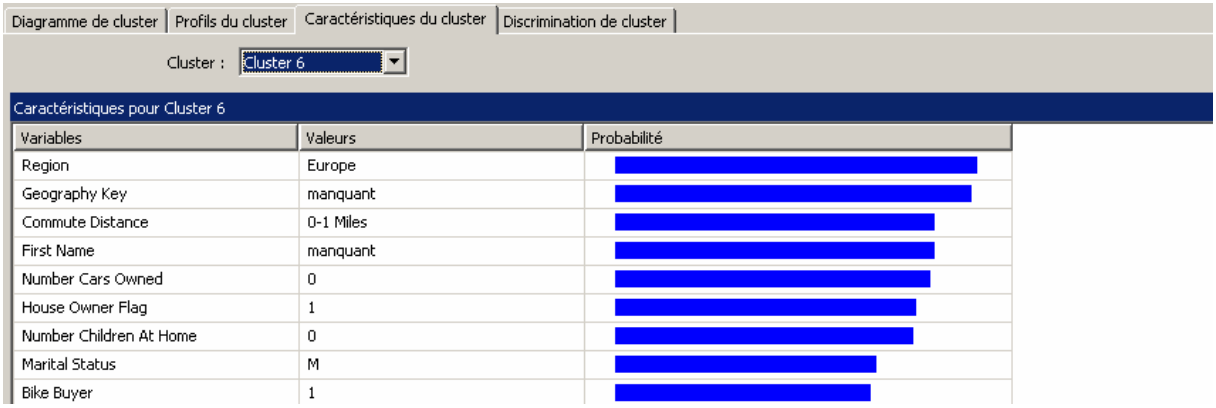


Figure 3-78 : Fenêtre « Publipostage.dmm » - Onglet « Visionneuse de modèles d'exploration de données » Caractéristique du cluster

3.6.1.4 Discrimination de cluster

Cet onglet (Figure 3-79) permet de comparer la valeur d'un attribut dans un cluster par rapport à un autre cluster.

Par exemple, nous choisissons comme « Cluster 1 : » : « Cluster 6 » et pour le « Cluster 2 : » : « Cluster 2 ».

Nous pouvons déduire, que dans le cluster 6, les cas gagnent entre 10'000.0 et 45'111.6 par années, alors dans le cluster 2 ils gagnent entre 45'111.6 et 170'000.0 annuellement..

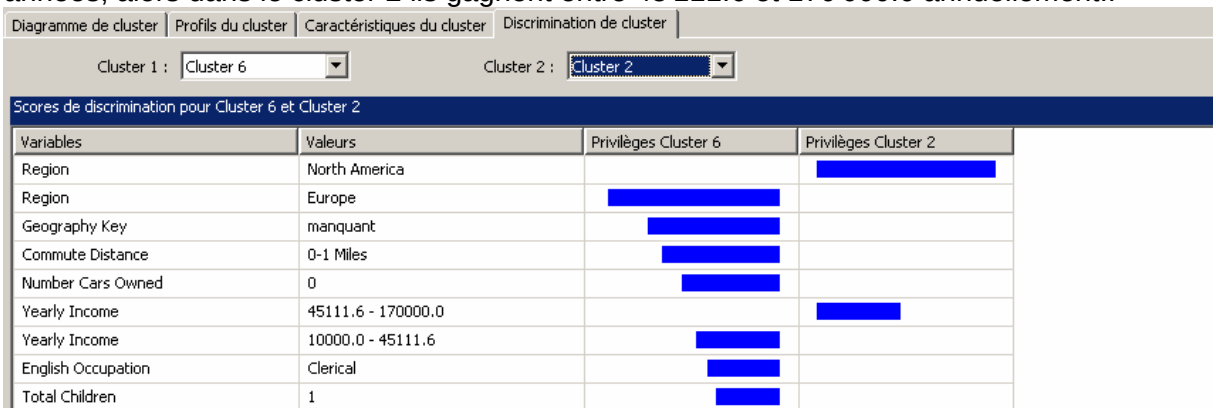


Figure 3-79 : Fenêtre « Publipostage.dmm » - Onglet « Visionneuse de modèles d'exploration de données » Discrimination de cluster

3.7 Graphique d'analyse de précision

Maintenant que nous avons préparé 3 modèles d'exploration de données, nous pouvons tester leurs différentes précisions pour effectuer de la prédiction et ainsi les comparer entre eux.
Nous choisissons donc le dernier onglet (« Graphique d'analyse de précision », Figure 3-80) que nous ne connaissons pas encore.

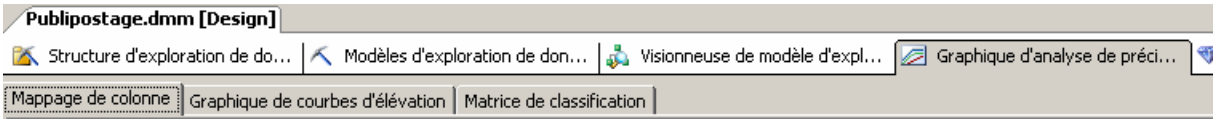


Figure 3-80 : Fenêtre « Publipostage.dmm » - Onglet « Graphique d'analyse de précision »

Dans cet onglet nous découvrons 3 sous onglets :

- Mappage de colonne
- Graphique de courbes d'élévation
- Matrice de classification

3.7.1 Mappage de colonne

L'onglet « Mappage de colonne » (Figure 3-81) permet de choisir quels sont les modèles d'exploration de données que nous désirons évaluer.

Nous commençons par sélectionner la table de cas, comme nous l'avons préalablement fait dans le chapitre « 3.4.5 Prévision de modèles d'exploration de données » à la différence que nous sélectionnons la table « vTargetMail » et non « ProspectiveBuyer ».
Nb. Nous reprenons la table fournie durant la phase d'apprentissage afin de réévaluer les cas qui ont servis lors de l'apprentissage.

Nous contrôlons aussi que tous les champs de la structure d'exploration de données soient mappés avec ceux de la table de cas et, dans le cas contraire, nous effectuons une jointure par glisser/déposer.

Dans la seconde partie de l'écran, nous pouvons appliquer des filtres pour exclure certains enregistrements du graphique ou, au contraire, forcer le graphique à afficher uniquement certains enregistrements.

La troisième partie de l'écran nous permet de choisir quels modèles d'exploration de données nous désirons tester ainsi, que dans le cas où il y en a plusieurs, quelle est la variable à prédire. Il est aussi de nécessaire d'indiquer à VS 2005 – BI quelle est la valeur que nous désirons prédire.

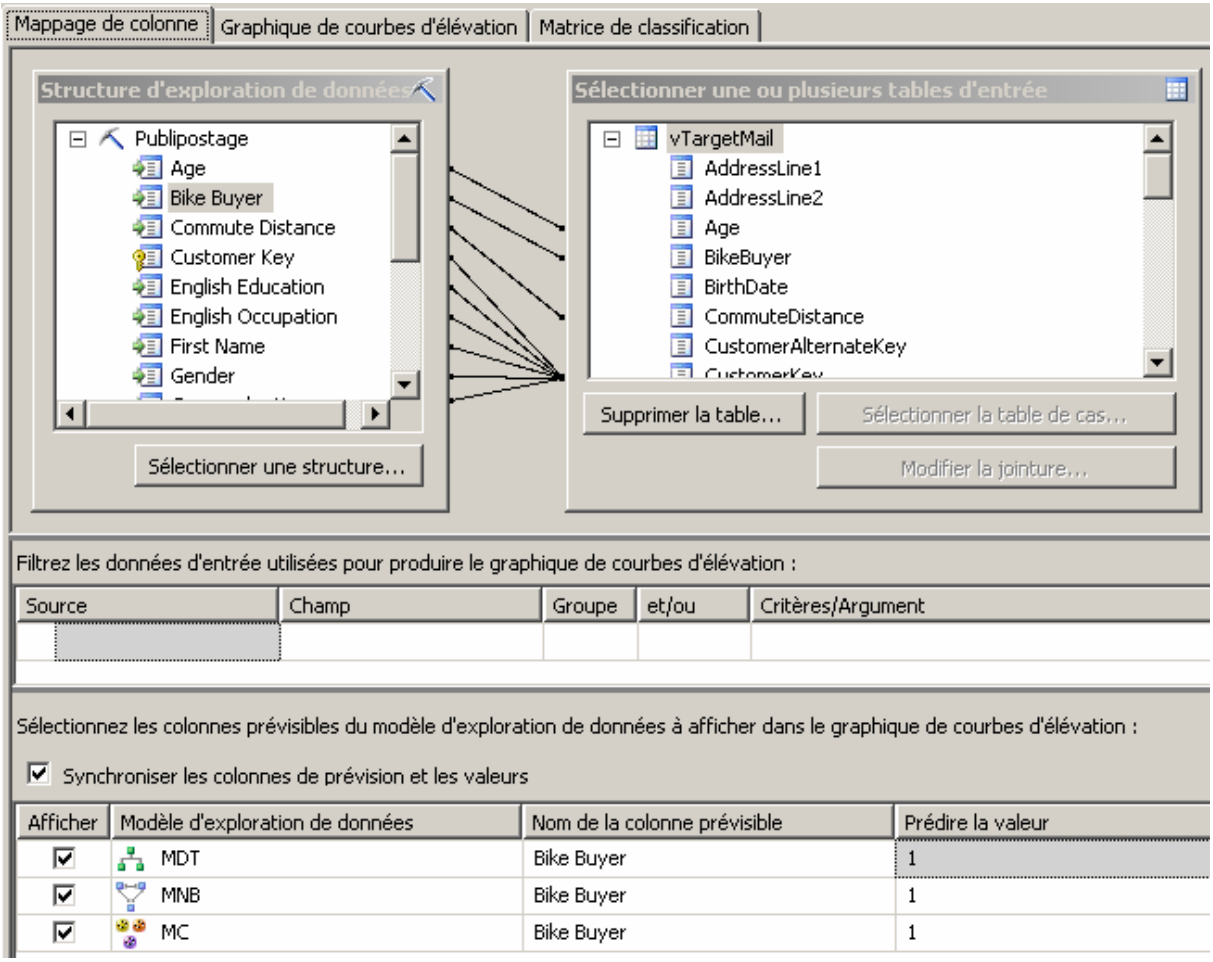


Figure 3-81 : Fenêtre « Publipostage.dmm » - Onglet « Graphique d'analyse de précision » - Mappage de colonne

3.7.2 Graphique de courbes d'élévation

Une fois les sélections effectuées dans l'onglet « Mappage de colonne », nous pouvons passer sur cet onglet (Figure 3-82) qui représente de manière graphique la précision de nos trois modèles d'exploration de données.

A l'aide de la « Légende d'exploration de données » (Figure 3-83), nous comprenons que la ligne verte du graphique indique la courbe d'élévation du modèle d'exploration MDT, la ligne violette celle du modèle MNB et la ligne jaune, représente la courbe du modèle MC. La courbe idéale est représentée en rouge.

Nous pouvons en déduire que le meilleur modèle d'exploration est le MDT.

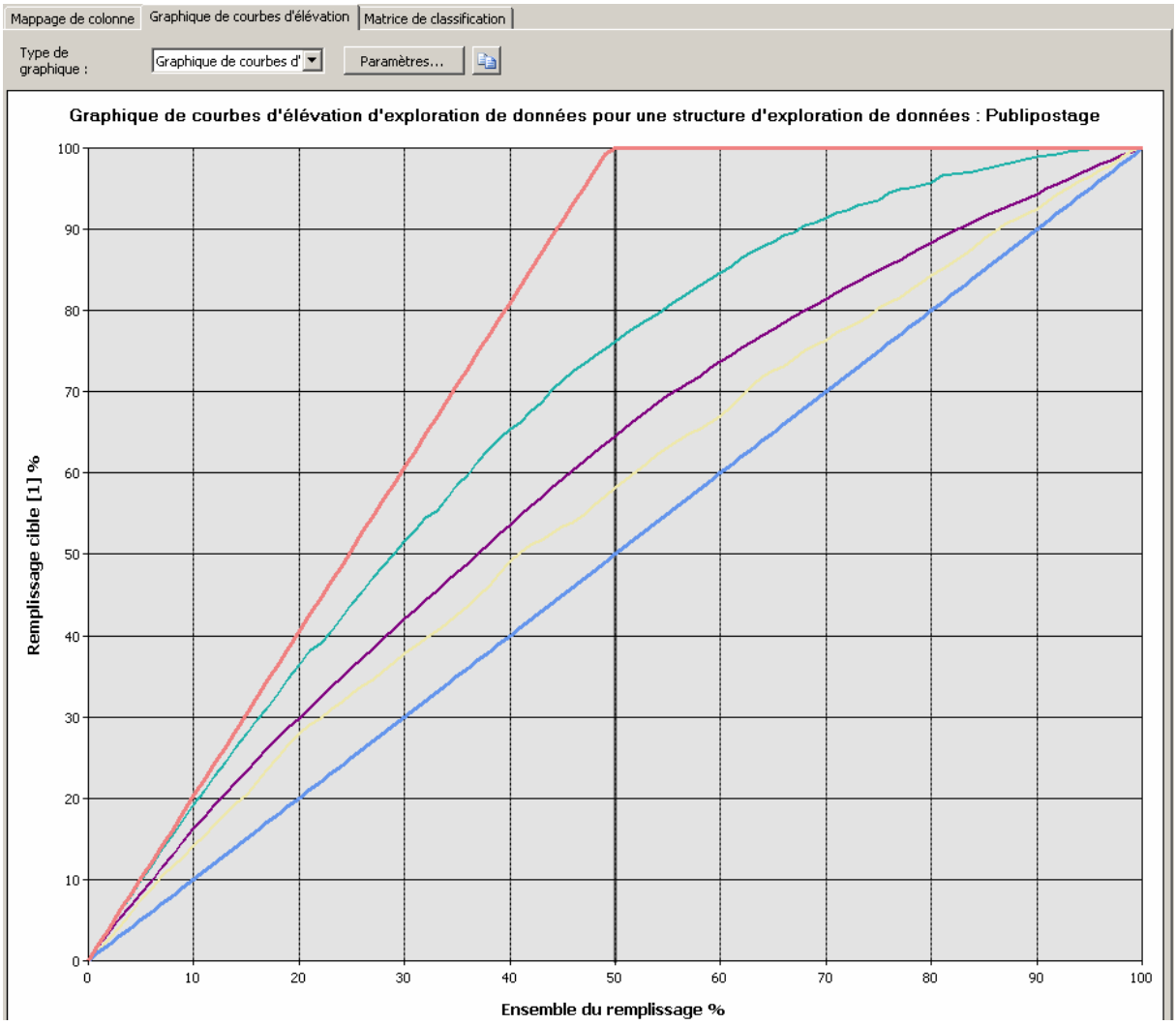


Figure 3-82 : Fenêtre « Publipostage.dmm » - Onglet « Graphique d'analyse de précision » - Graphique de courbes d'élévation

Légende d'exploration de données

Pourcentage de remplissage : 50.00%

Série, modèle	Score	Remplissage cible	Prédire la probabilité
MDT	0.89	76.19%	45.25%
MNB	0.79	64.48%	50.15%
MC	0.74	58.21%	47.12%
Modèle d'estimation aléatoire		50.00%	
Modèle idéal pour : MDT, MNB, MC		100.00%	

Figure 3-83 : Fenêtre « Publipostage.dmm » - Onglet « Graphique d'analyse de précision » - Graphique de courbes d'élévation - Légende d'exploration de données

A partir de la liste déroulante « Type de graphique : » au sommet à gauche de l'onglet « Graphique de courbes d'élévation » nous pouvons créer un graphique de bénéfice afin de comparer les différents modèles et les bénéfices qu'ils pourraient engendrer.

Lors du choix de la liste déroulante pour sélectionner un graphique des bénéfices, VS2005 – BI affiche la fenêtre « Paramètre du graphique des bénéfices » (Figure 3-84) afin que nous lui indiquions quelques éléments financiers. Nous reprenons les indications figurant dans la Figure 3-84.

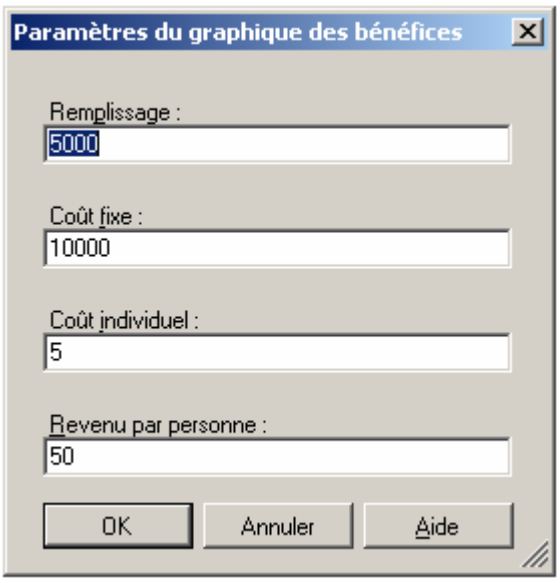


Figure 3-84 : Fenêtre « Publipostage.dmm » - Onglet « Graphique d'analyse de précision » - Graphique de courbes d'élévation – Paramètres du graphique des bénéfices

Une fois les informations saisies, nous cliquons sur le bouton « OK » afin que VS2005 – BI nous affiche le résultat demandé dans un graphique (Figure 3-85)

Nous pouvons à nouveau observer que le modèle d'exploration MDT propose une meilleure courbe que les deux autres modèles.

La « légende d'exploration de données » (Figure 3-86) permet de distinguer dans le graphique les différents modèles d'exploration de données et indique aussi le bénéfice réalisé par chacun des modèles.

De plus, dans le graphique des bénéfices (Figure 3-85) nous pouvons déplacer une droite (qui se trouve ici à 50%) afin d'évaluer les bénéfices effectués à un moment x.

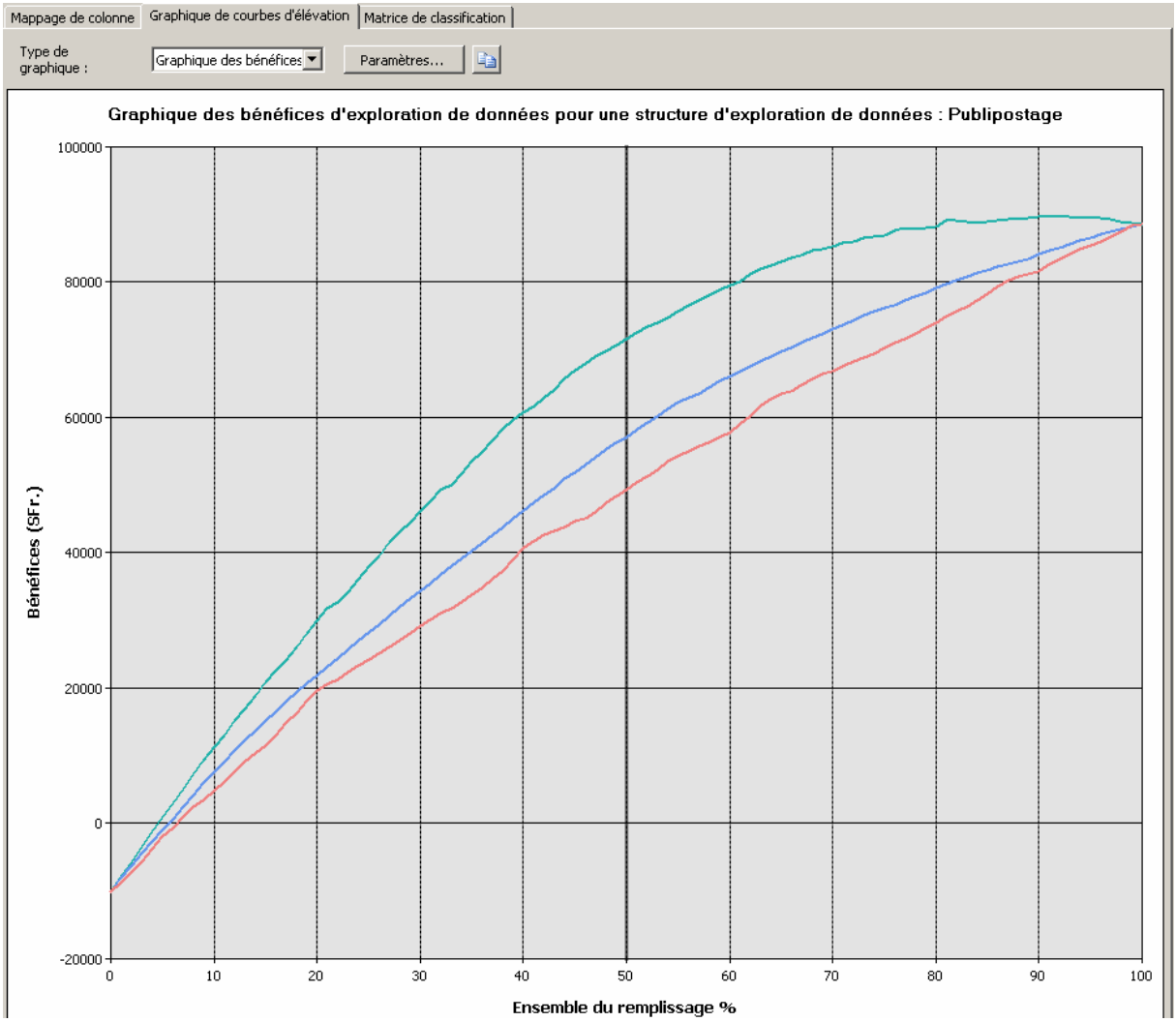


Figure 3-85 : Fenêtre « Publipostage.dmm » - Onglet « Graphique d'analyse de précision » - Graphique de courbes d'élévation – Avec bénéfice

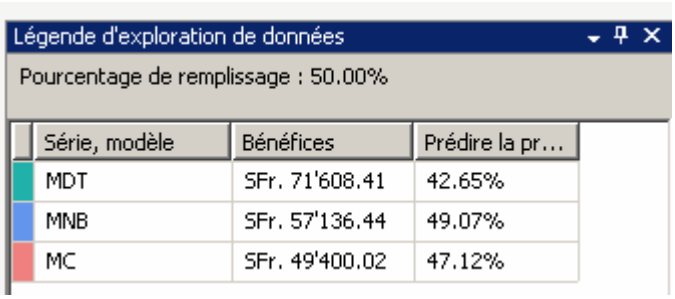


Figure 3-86 : Fenêtre « Publipostage.dmm » - Onglet « Graphique d'analyse de précision » - Graphique de courbes d'élévation - Avec bénéfice - Légende d'exploration de données

3.7.3 Matrice de classification

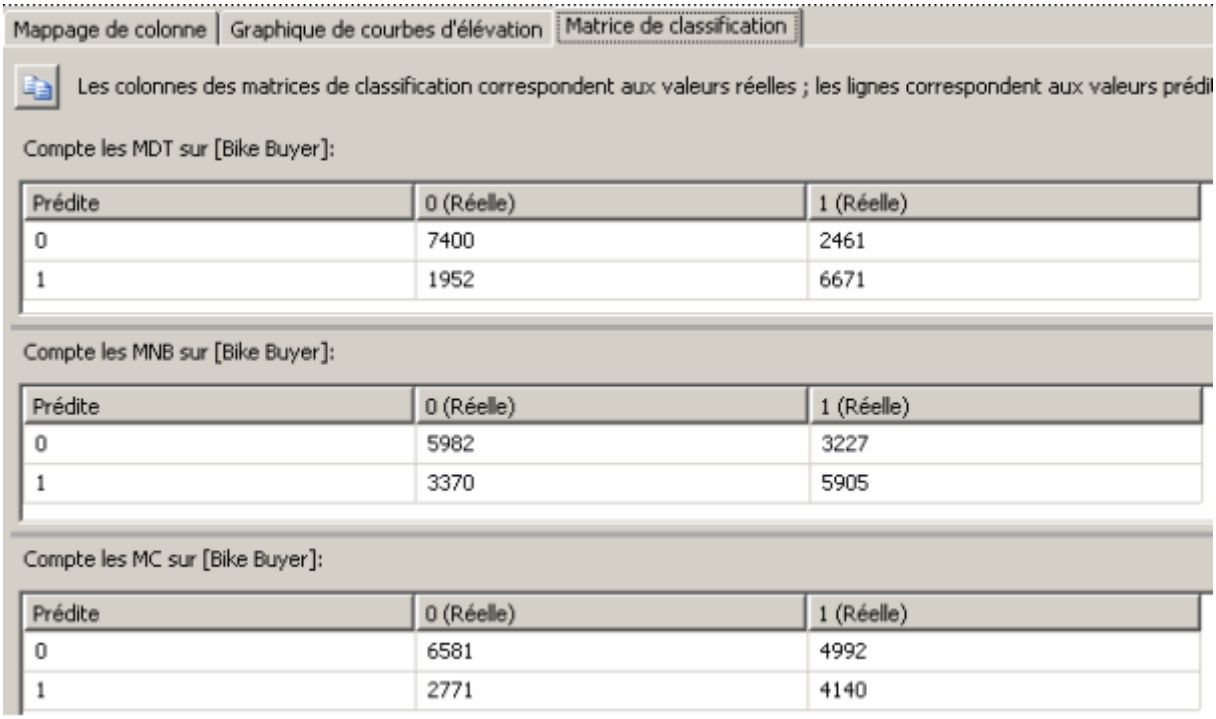
Dans cet onglet (Figure 3-87), VS2005 – BI compare les diverses prédictions effectuées par rapport aux valeurs contenues dans la table d'apprentissage.

Nous pouvons y observer que pour le modèle d'exploration « MDT », VS2005 – BI nous a prédit correctement 7400 « Bike Buyer = 0 » (40% des cas totaux) et 6671 « Bike Buyer = 1 » (36% des cas totaux). Ce qui signifie qu'il n'a pas prédit correctement 4413 acheteur (+/- 24% des cas totaux).

Si nous extrapolons cette déduction pour les deux autres modèles d'exploration de données, nous obtenons :

- pour le modèle MNB
  5982 « Bike Buyer = 0 » (32% des cas totaux)
  5905 « Bike Buyer = 1 » (32% des cas totaux)
  6597 erreurs (+/- 36% des cas totaux)
- pour le modèle MC
  6581 « Bike Buyer = 0 » (37% des cas totaux)
  4140 « Bike Buyer = 1 » (23% des cas totaux)
  7263 erreurs (+/- 40% des cas totaux)

Cet onglet nous permet d'affirmer nos conclusions effectuées à partir du « Graphique de courbes d'élévation » qui sont que le meilleur modèle de prédiction est le modèle d'exploration de données « MDT ».


Mappage de colonne | Graphique de courbes d'élévation | Matrice de classification

Les colonnes des matrices de classification correspondent aux valeurs réelles ; les lignes correspondent aux valeurs prédi

Compte les MDT sur [Bike Buyer]:

Prédite	0 (Réelle)	1 (Réelle)
0	7400	2461
1	1952	6671

Compte les MNB sur [Bike Buyer]:

Prédite	0 (Réelle)	1 (Réelle)
0	5982	3227
1	3370	5905

Compte les MC sur [Bike Buyer]:

Prédite	0 (Réelle)	1 (Réelle)
0	6581	4992
1	2771	4140

Figure 3-87 : Fenêtre « Publipostage.dmm » - Onglet « Graphique d'analyse de précision » - Matrice de classification

4 Microsoft SQL Server Management Studio

Voici une brève description de SQL Server 2005 Analysis Services (SSAS) et de l'interface Microsoft SQL Server Management Studio (SSMS).

Cette interface est moins conviviale que celle de Microsoft Visual Studio 2005 – Business Intelligence, mais possède un avantage certain qui est le fait de ne pas avoir besoin d'installer l'environnement Visual Studio 2005 : seul le client Microsoft SQL Server Management Studio suffit.

Pour se lancer avec SSMS, nous exécutons à partir du menu « Démarrer, Programmes, Microsoft SQL Server 2005 » le programme « SQL Server Management Studio ».

La fenêtre « Se connecter au serveur » (Figure 4-1) s'affiche. Dans cet écran, nous sélectionnons, dans la liste déroulante correspondante, le type de serveur « Analysis Services ». Ensuite, nous choisissons, dans la liste déroulante suivante le nom du serveur (l'instance) auquel nous voulons nous connecter (ici « TESTSERVER ») et nous cliquons sur « OK ».

Figure 4-1 : Se connecter au serveur

Une fois connecté au serveur d'analyse (Figure 4-2), nous pouvons observer, sous le nom du serveur, deux répertoires :

- Bases de données
  Dans SSMS, une base de données correspond au nom du projet dans VS2005 – BI.
- Assemblys

Nb. Dans ce tutorial, nous nous intéressons uniquement au répertoire « Bases de données ».

Nous développons donc le dossier « Base de données ».

Etant donné que nous avons déjà créé dans le chapitre « 3 - Microsoft Visual Studio 2005 – BI – tutorial » une base de données de Data Mining, celle-ci y figure.
Nous pouvons, et nous le faisons, créer une nouvelle base de données en effectuant un clic droit sur le dossier « Bases de données » et nous choisissons, dans le menu contextuel, « Nouvelle base de données ».

Dans l'écran « Nouvelle base de données » (Figure 4-3), nous donnons un nom à cette nouvelle base de données (ici « Tutorial Data Mining SSMS »), nous sélectionnons l'option

« Utiliser le compte de service » pour l'« Emprunt d'identité » et nous validons le tout en cliquant sur « OK ».

Le dossier « Bases de données » contient maintenant les deux bases de données.

A partir de ce point, la souris devient presque totalement inutile car toutes les futures opérations se font en langages DMX (du langage SQL adapté au monde du Data Mining).

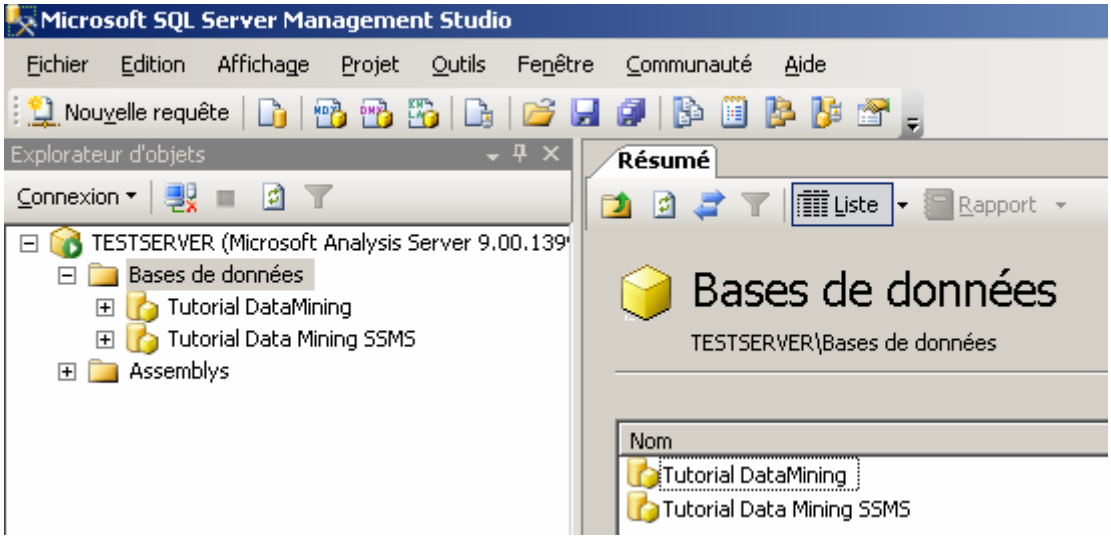


Figure 4-2 : Microsoft SQL Server Management Studio

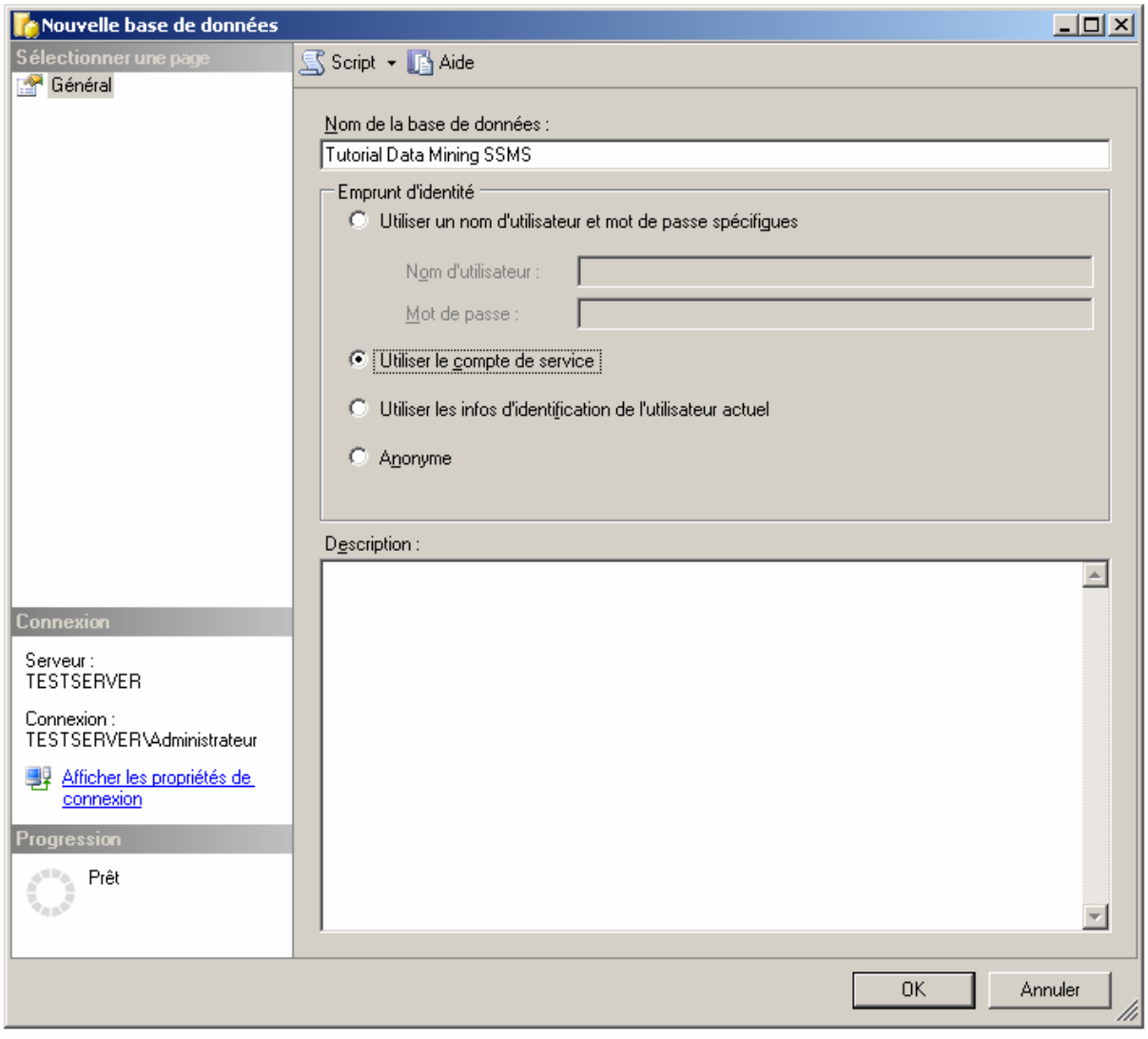


Figure 4-3 : Nouvelle base de données

4.1 Le langage DMX

Afin de faire une approche simplifiée de ce langage, nous refaisons le même scénario de Data Mining que celui abordé dans le chapitre « 3 - Microsoft Visual Studio 2005 – BI – tutorial», mais en créant uniquement le modèle d'exploration de données « Microsoft Decision Trees ».

Pour créer une requête DMX, nous cliquons sur le bouton « Requête DMX Analysis Services » ( ) situé dans la barre d'outils « Standard » de SSMS.
Lors du clic de ce bouton SSMS affiche à nouveau la Figure 4-1 pour nous inviter à connecter l'éditeur de requête directement au serveur SSAS sur lequel nous désirons effectuer les requêtes.
De plus, SSMS affiche la barre d'outils « Editeur SQL Server Analysis Services » (Figure 4-4).

Figure 4-4 : Barre d'outil « Editeur SQL Serveur Analysis Services »

Par défaut, SSMS affiche un modèle d'exploration de données pour la base de données affichée dans la liste déroulante de la Figure 4-4 (ici « Tutorial DataMining ».

Bien que nous n'utilisions pas le volet de gauche de l'éditeur de requête, nous allons le parcourir rapidement. Nous pouvons apercevoir que ce volet contient deux onglets :

- Métadonnées
- Fonctions

L'onglet « Métadonnées » (Figure 4-5 Figure 4-6) contient les attributs du modèle d'exploration sélectionné. Nous pouvons distinguer la clé (« Customer Key »), le/les attributs à prédire (« Bike Buyer ») et les attributs d'entrées ( ).

L'onglet « Fonctions » (Figure 4-5 Figure 4-6) contient toutes les fonctions DMX des divers algorithmes SSAS ainsi que d'autres fonctions relatives à l'élaboration et à la navigation concernant le Data Warehousing.

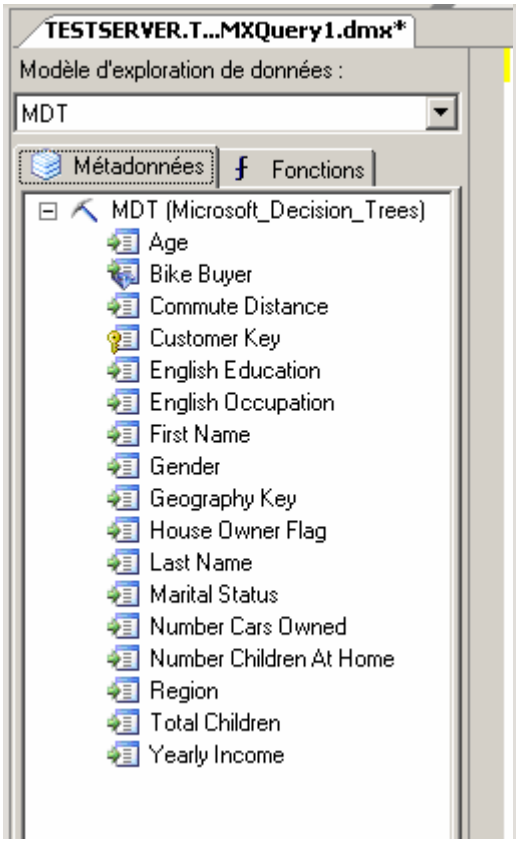


Figure 4-5

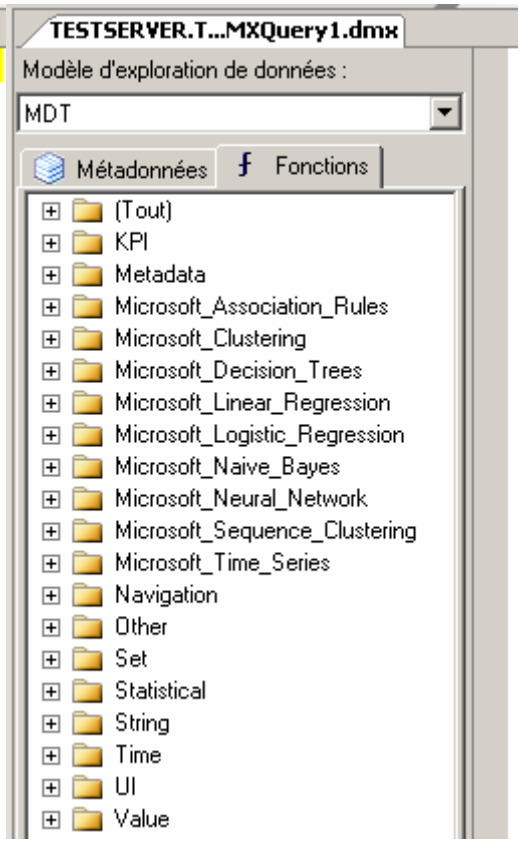


Figure 4-6¨

Etant donné que notre nouvelle base de données ne contient actuellement aucun modèle d'exploration, les deux volets de l'éditeur de requête affiche un message d'erreur (Figure 4-7 et Figure 4-8).

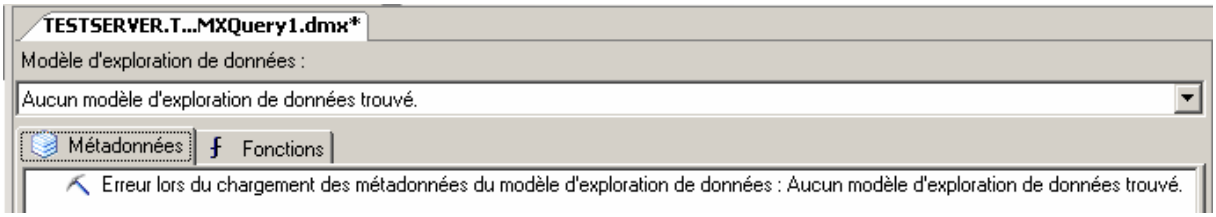


Figure 4-7 : Fenêtre « MXQuery1.dmx » - Onglet « Métadonnées »

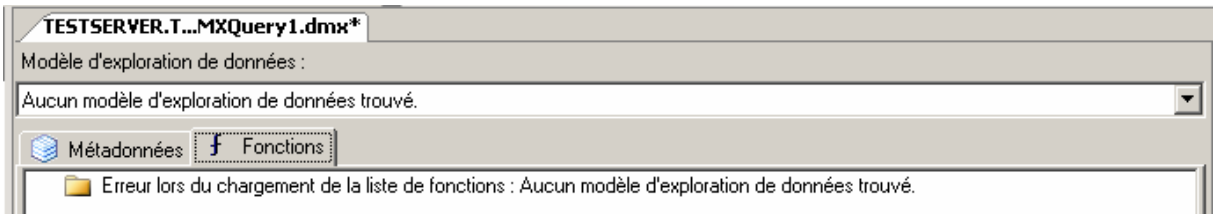


Figure 4-8 : Fenêtre « MXQuery1.dmx » - Onglet « Fonctions »

4.1.1 Créer une Structures d'exploration de données

Par rapport à un projet VS2005 – BI, dans SSMS nous n'avons pas besoin de créer une « Source de données », ni une « Vue des sources de données ».

Pour créer une « Structure d'exploration de données », nous utilisons l'instruction DMX « `CREATE MINING STRUCTURE <NomStructure>` » (Script 4-1):

```
CREATE MINING STRUCTURE [Publipostage_SSMS]
(
	[Customer Key] LONG KEY,
	[Age] LONG DISCRETIZED,
	[Commute Distance] TEXT DISCRETE,
	[English Education] TEXT DISCRETE,
	[English Occupation] TEXT DISCRETE,
	[First Name] TEXT DISCRETE,
	[Gender] TEXT DISCRETE,
	[Geography Key] LONG DISCRETE,
	[House Owner Flag] TEXT DISCRETE,
	[Last Name] TEXT DISCRETE,
	[Marital Status] TEXT DISCRETE,
	[Number Cars Owned] LONG DISCRETE,
	[Number Children At Home] LONG DISCRETE,
	[Region] TEXT DISCRETE,
	[Total Children] LONG DISCRETE,
	[Yearly Income] DOUBLE CONTINUOUS,
	[Bike Buyer] LONG DISCRETE
)
```

Script 4-1

Une fois l'instruction saisie, nous cliquons sur le bouton « Exécuter (!) » de la barre d'outils « Editeur SQL Server Analysis Services » afin d'exécuter la requête.

Une fois la requête exécutée, le volet « Message » (Figure 4-9) affiche le message « Exécution terminée » indiquant le bon déroulement de l'exécution.

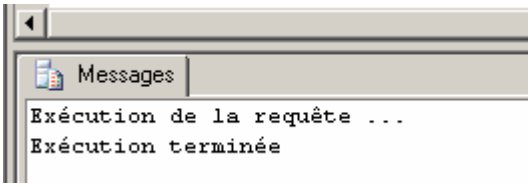


Figure 4-9 : Fenêtre « MXQuery1.dmx » - Onglet « Message » - Exécution de la requête

Nous pouvons développer le répertoire « Bases de données » de l'« Explorateur d'objet » jusqu'à atteindre notre nouveau modèle d'exploration se trouvant dans le dossier « Structures d'exploration de données » de la base de données « Tutorial Data Mining SSMS » (Figure 4-10).

Si nous descendons encore d'un dossier, nous apercevons que le répertoire « Modèle d'exploration de données » est encore vide.
Nous remédions à ce problème au prochain chapitre.

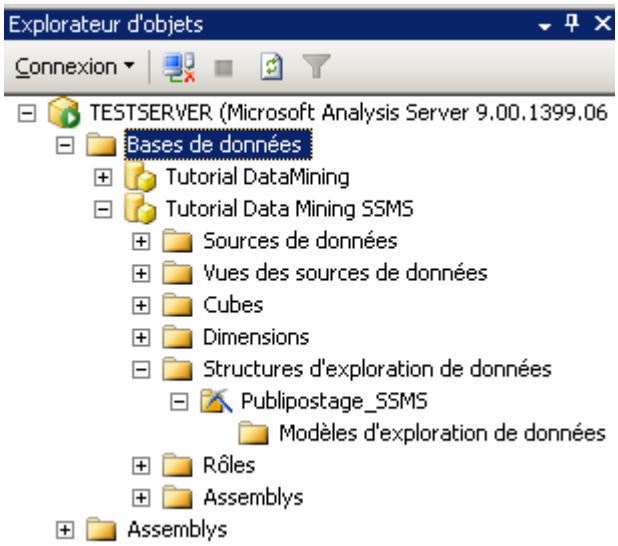


Figure 4-10 : Fenêtre « Explorateur d'objets »

4.1.2 Créer un Modèle d'exploration de données

L'instruction DMX qui permet d'ajouter un modèle d'exploration de données à une structure d'exploration de données est « `ALTER MINING STRUCTURE <NomStructure> ADD MINING MODEL <NomModele>` ».
Nous saisissons l'instruction suivante (Script 4-2) afin d'ajouter à la structure d'exploration de données « Publipostage_SSMS » un modèle d'exploration de données basé sur l'algorithme « Microsoft Decision Trees » que nous nommons MDT_SSMS.

```
ALTER MINING STRUCTURE [Publipostage_SSMS]
ADD MINING MODEL MDT_SSMS
(
      [Customer Key],
      [Age],
      [Commute Distance],
      [English Education],
      [English Occupation],
      [First Name],
      [Gender],
      [Geography Key],
      [House Owner Flag],
      [Last Name],
      [Marital Status],
      [Number Cars Owned],
      [Number Children At Home],
      [Region],
      [Total Children],
      [Yearly Income] REGRESSOR,
      [Bike Buyer] PREDICT_ONLY
)
USING MICROSOFT_DECISION_TREES
```

Script 4-2

Une brève explication de ce script s'avère utile :

- lorsque le nom du champ n'est pas suivi d'un mot clé, cela signifie que ce champ est défini en tant qu'attribut d'entrée pour l'algorithme.
- le mot clé « `REGRESSOR` » est à utiliser lorsque nous avons indiqué à la structure d'exploration de données un champ « `CONTINUOUS` ».
- les mots clé « `PREDICT` » et « `PREDICT_ONLY` » signifie que ce champ est l'attribut que nous souhaitons prédire.

4.1.3 Traiter un Modèle d'exploration de données

L'instruction DMX qui permet de traiter un modèle d'exploration de données correspond à « `INSERT INTO MINING STRUCTURE <NomStructure>` ».

Pour se connecter à la base de données qui contient les données d'apprentissage, nous pouvons utiliser deux méthodes :

- `OPENQUERY`
- `OPENROWSET`

La méthode « `OPENQUERY` » peut être utilisée UNIQUEMENT si nous nous connectons à une base de donnée Data Mining qui est crée via VS2005 – BI, car nous devons y spécifier une « Source de données ». Le problème est qu'avec le langage DMX nous ne pouvons pas créer une « Source de données » et que l'environnement SSMS ne le permet pas non plus…

L'exemple ci-dessous (Script 4-3) simule le fait que nous nous soyons connectés sur la base de données de Data Mining créée au chapitre « 3 - Microsoft Visual Studio 2005 – BI – tutorial » et que nous retraitons le modèle d'exploration MDT.

```
INSERT INTO MDT
(
      [Customer Key],                [Age],
      [Commute Distance],            [English Education],
      [English Occupation],          [First Name],
      [Gender],                      [Geography Key],
      [House Owner Flag],            [Last Name],
      [Marital Status],              [Number Cars Owned],
      [Number Children At Home],     [Region],
      [Total Children],              [Yearly Income],
      [Bike Buyer]
)
OPENQUERY([Adventure Works DW],
'SELECT
      CustomerKey,                   Age,
      CommuteDistance,               EnglishEducation,
      EnglishOccupation,             FirstName,
      Gender,                        GeographyKey,
      HouseOwnerFlag,                LastName,
      aritalStatus,                  umberCarsOwned,
      NumberChildrenAtHome,          egion,
      TotalChildren,                 YearlyIncome,
      BikeBuyer
FROM  AdventureWorksDW.dbo.vTargetMail')
```

Script 4-3

Afin d'utiliser l'instruction « OPENROWSET» (la seule méthode disponible dans un projet SSAS totalement SSMS), il est nécessaire de modifier la configuration du serveur SSAS pour autoriser des requêtes AD HOC et spécifier quels sont les « Provider » autorisés à effectuer ces requêtes.

Pour accéder au panneau de contrôle de SSAS, nous faisons un clic droit sur le nom du serveur depuis l'« Explorateur d'objet » et nous choisissons dans le menu contextuel « Propriétés » (Figure 4-11).

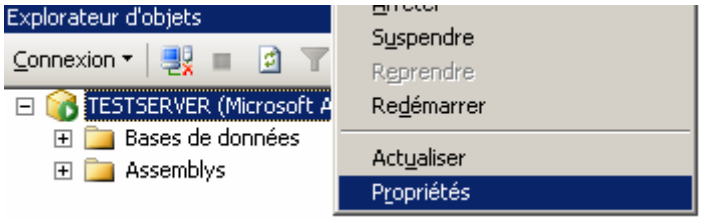


Figure 4-11 : Fenêtre « Explorateur d'objets » - Menu contextuel

L'écran « Propriétés de Analysis Server » (Figure 4-12) s'affiche. Depuis cet écran, nous commençons par changer la valeur du champ :
« DataMining \ AllowAdHocOpenRowsetQueries » de « False » à « True » (en surbrillance dans la Figure 4-12).

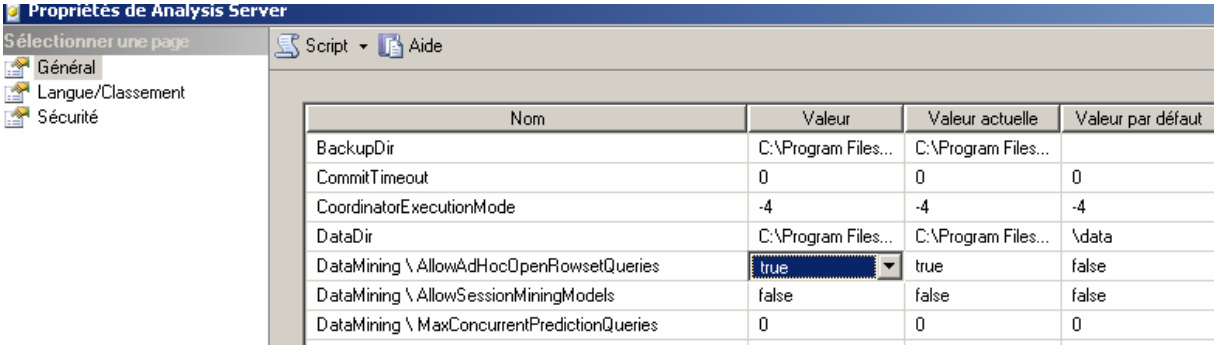
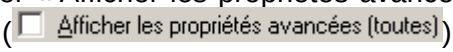


Figure 4-12 : Fenêtre « Propriété de Analysis Server » - Option « AllowAdHocRowsetQueries »

Ensuite, nous devons activer la case à cocher « Afficher les propriétés avancées (toutes) » se situant en bas à gauche de la Figure 4-12 (

).

Toutes les propriétés de SSAS s'affichent et parmi tous ces paramètres, il faut ajouter à celui qui se nomme « DataMining \ AllowedProvidersInOpenRowser » la valeur « SQLOLEDB » (en surbrillance dans la Figure 4-13).

DataMining \ Algorithms \ Microsoft_Sequence_Clustering \ Enabled	true	true
DataMining \ Algorithms \ Microsoft_Time_Series \ Enabled	true	true
DataMining \ AllowAdHocOpenRowsetQueries	true	true
DataMining \ AllowedProvidersInOpenRowset	SQLOLEDB	SQLOLEDB
DataMining \ AllowSessionMiningModels	false	false
DataMining \ MaxConcurrentPredictionQueries	0	0
DataMining \ Services \ MicrosoftAssociationRules \ MaximumItemsetCount	200000	200000
DataMining \ Services \ MicrosoftAssociationRules \ MaximumItemsetSize	3	3

Figure 4-13 : Fenêtre « Propriété de Analysis Server » - Option « AllowedProvidersInOpenRowset »

Nous validons ces deux changements en cliquant sur « OK ».

La première partie de l'instruction DMX est pareille à celle décrite plus haut. Par contre, la partie délicate de cette instruction, mis à part la reconfiguration de SSAS, est le fait que nous devions nous connecter à une base de données via le script DMX et que le mot de passe y est affiché en clair.

Bien entendu, il est possible de créer un utilisateur de base de donnée ayant uniquement les droits de lecture sur la table source et rien d'autre, limitant ainsi les risques.

Nous exécutons donc ce script (Script 4-4) sur la base de donnée de Data Mining « Tutorial Data Mining SSMS ».

Dès la fin de l'exécution du script, le volet de gauche de l'éditeur de requête nous affiche, sous l'onglet « Métadonnées », le modèle d'exploration de données nouvellement créé.

```
INSERT INTO MDT_SSMS
(
      [Customer Key],               [Age],
      [Commute Distance],           [English Education],
      [English Occupation],         [First Name],
      [Gender],                     [Geography Key],
      [House Owner Flag],           [Last Name],
      [Marital Status],             [Number Cars Owned],
      [Number Children At Home],    [Region],
      [Total Children],             [Yearly Income],
      [Bike Buyer]
)
OPENROWSET('SQLOLEDB',
'Provider=SQLOLEDB;Persist Security
Info=False;Trusted_Connection=yes;Initial Catalog=AdventureWorksDW;Data
Source=TESTSERVER;UID=SSAS;PWD=ssas1',
'SELECT
      CustomerKey,                  Age,
      CommuteDistance,              EnglishEducation,
      EnglishOccupation,            FirstName,
      Gender,                       GeographyKey,
      HouseOwnerFlag,               LastName,
      aritalStatus,                 umberCarsOwned,
      NumberChildrenAtHome,         egion,
      TotalChildren,                YearlyIncome,
      BikeBuyer
FROM  AdventureWorksDW.dbo.vTargetMail')
```

Script 4-4

4.1.4 Exécution d'une requête de prédiction

Pour générer une requête de prédiction dans SSMS, nous avons deux possibilités selon la manière dont a été créée la base de données de Data Mining.

Si nous naviguons dans une base de données de Data Mining créer avec VS2005 – BI, nous pouvons simplement préparer une requête de prédiction via le même outil graphique du chapitre « 3.4.5 - Prévision de modèles d'exploration de données» en faisant un clic droit sur le modèle d'exploration de données voulus (ici « MDT») (Figure 4-14) et en choisissant dans le menu contextuel l'option « Générer une requête de prédiction… ».

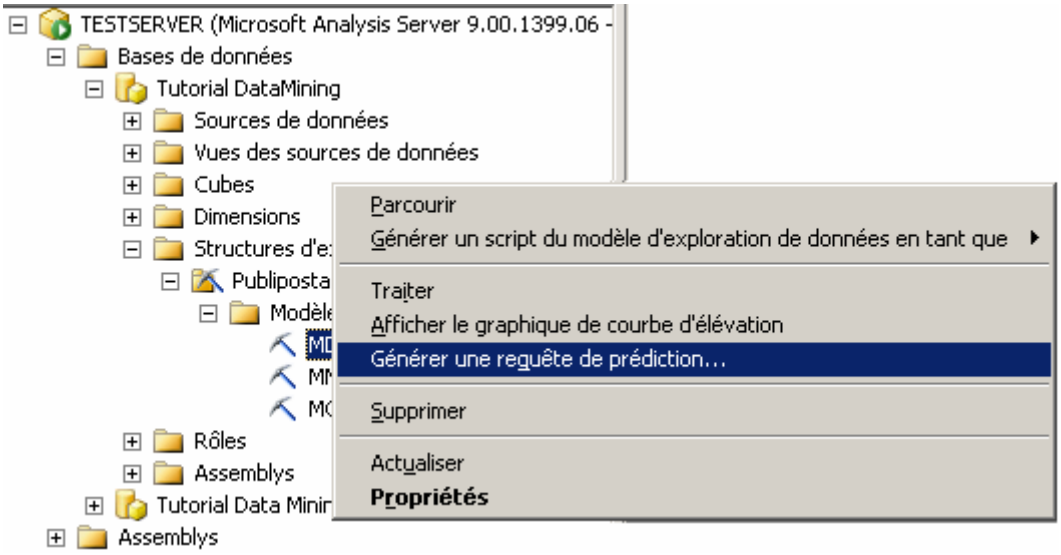


Figure 4-14 : Fenêtre « Explorateur d'objets » - Menu Contextuel MDT

Une fois l'outil graphique à l'écran, celui-ci s'utilise exactement comme celui du chapitre « 3.4.5 - Prévision de modèles d'exploration de données ».

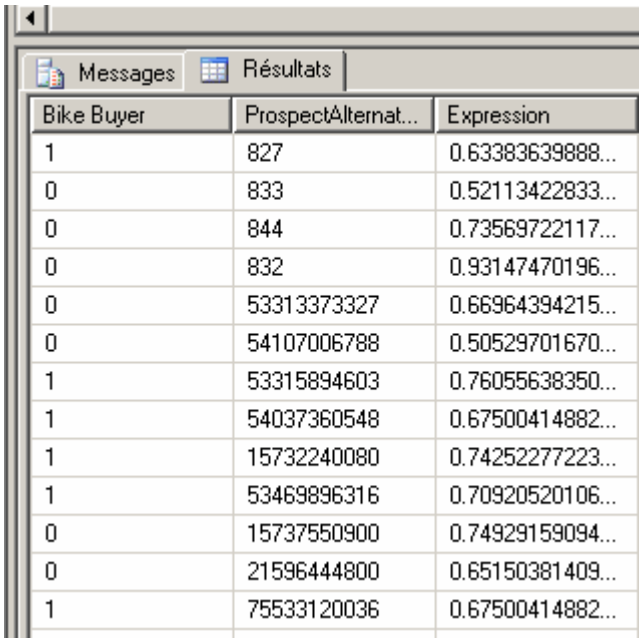
Mais étant donné que ce chapitre est consacré à l'utilisation de SSMS et des requêtes DMX, nous préparons directement le script DMX suivant (Script 4-5) :

```
SELECT
  [MDT_SSMS].[Bike Buyer],
  t.[ProspectAlternateKey],
  PredictProbability([MDT_SSMS].[Bike Buyer])
FROM
  [MDT_SSMS]
PREDICTION JOIN
OPENROWSET('SQLOLEDB',
'Provider=SQLOLEDB;Persist Security
Info=False;Trusted_Connection=yes;Initial Catalog=AdventureWorksDW;Data
Source=TESTSERVER;UID=SSAS;PWD=ssas1',
    'SELECT
      [ProspectAlternateKey],
      [MaritalStatus],
      [Gender],
      [YearlyIncome],
      [TotalChildren],
      [NumberChildrenAtHome],
      [HouseOwnerFlag],
      [NumberCarsOwned],
      [Age],
      [Education],
      [Occupation]
    FROM
      [dbo].[ProspectiveBuyer]
    ') AS t
ON
  [MDT_SSMS].[Marital Status] = t.[MaritalStatus] AND
```

```
  [MDT_SSMS].[Gender] = t.[Gender] AND
  [MDT_SSMS].[Yearly Income] = t.[YearlyIncome] AND
  [MDT_SSMS].[Total Children] = t.[TotalChildren] AND
  [MDT_SSMS].[Number Children At Home] = t.[NumberChildrenAtHome] AND
  [MDT_SSMS].[House Owner Flag] = t.[HouseOwnerFlag] AND
  [MDT_SSMS].[Number Cars Owned] = t.[NumberCarsOwned] AND
  [MDT_SSMS].[Age] = t.[Age] AND
  [MDT_SSMS].[English Education] = t.[Education] AND
  [MDT_SSMS].[English Occupation] = t.[Occupation]
```

Script 4-5

Les résultats de cette requête sont affichés dans le volet du bas de l'éditeur de requête, dans l'onglet « Résultat » (Figure 4-15).



Bike Buyer	ProspectAlternat...	Expression
1	827	0.63383639888...
0	833	0.52113422833...
0	844	0.73569722117...
0	832	0.93147470196...
0	53313373327	0.66964394215...
0	54107006788	0.50529701670...
1	53315894603	0.76055638350...
1	54037360548	0.67500414882...
1	15732240080	0.74252277223...
1	53469896316	0.70920520106...
0	15737550900	0.74929159094...
0	21596444800	0.65150381409...
1	75533120036	0.67500414882...

Figure 4-15 : Fenêtre « MXQuery1.dmx » - Onglet « Résultats »

Il va de soi que les résultats retournés par l'une ou l'autre méthodes seront égaux.

5 Conclusion

Durant ce tutorial, nous avons testé différentes méthodes pour l'exploration et l'exploitation de données.

Bien que la partie VS2005 – BI soit plus convivial à utiliser, lorsque nous voulons faire du Data Mining à partir d'autres logiciels (MS Excel, page Web…, il est nécessaire de maîtriser au minimum le langage DMX qui permet d'effectuer des prédictions.
Par exemple, dans un site web ASP, il est nécessaire de créer la requête DMX et, dans MS Excel, seule la version 2007 permet de se connecter à un serveur d'analyse.

Pour une mise en production d'un serveur d'analyse, je conseille de commencer le projet à l'aide de VS2005 – BI, qui permet de réaliser des modèles d'exploration de données beaucoup plus rapidement et, ensuite, l'implémentation peut se faire à l'aide du langage DMX.
De plus, en passant par un projet VS2005 – BI, nous pouvons créer des « Sources de données » ainsi que des « Vues de source de données », ce qui permet de masquer, dans les requête DMX, les propriétés de connexion, offrant ainsi plus de sécurité.

6 Annexes

6.1 Installation des bases de données de tests Microsoft

Pour installer les bases de test Microsoft SQL Serveur 2005, il faut passer par les menus : Démarrer, Paramètres, Panneau de configuration, double cliquer sur l'icône « Ajout/Suppression de programmes », sélectionner dans la liste des programmes installés proposés : « Microsoft SQL Server 2005 » et cliquer sur le bouton « Modifier ».

La fenêtre « Maintenance de Microsoft SQL Server 2005 » (Figure 6-1) s'affiche. Dans celle-ci, nous choisissons l'option : « Composants de la station de travail » et nous cliquons sur « Suivant ».

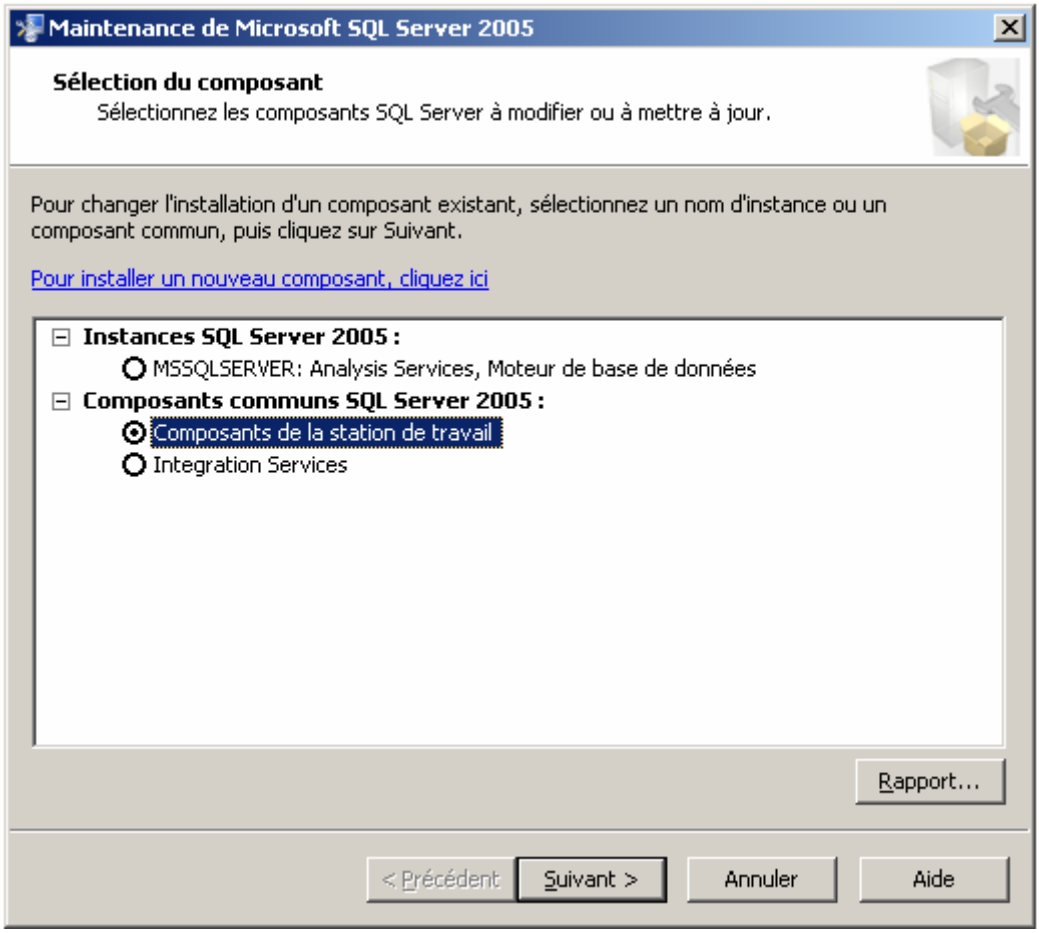


Figure 6-1 : Maintenance de Microsoft SQL Server 2005

A la suite de cet écran, deux fenêtres se succèderont. Dans chacune d'elles, nous cliquons sur le bouton « Suivant ».

Ensuite, à l'écran « Changer ou supprimer l'instance » (Figure 6-2), nous cliquons sur le bouton « Changer les composants installés » et nous validons en cliquant sur Suivant.

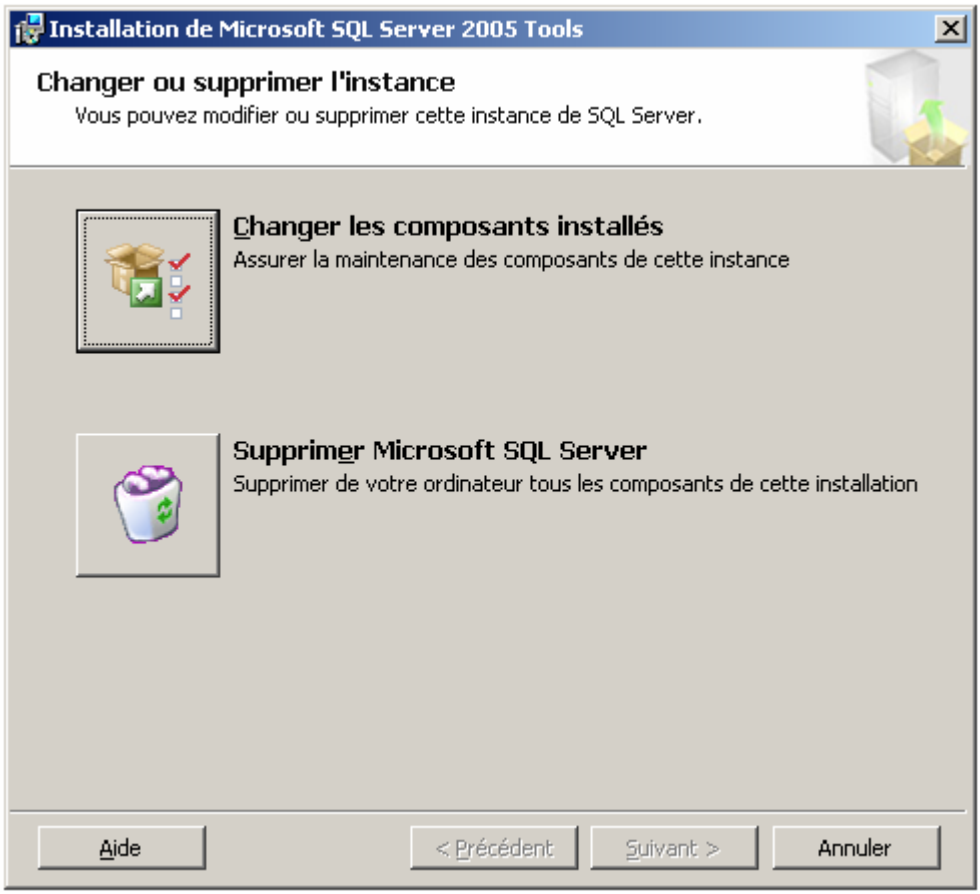


Figure 6-2 : Changer ou supprimer une instance

A l'écran « Sélection de composant » (Figure 6-3), nous développons le nœud « Documentation, exemples et exemples de bases de données » jusqu'à atteindre le nœud final « Exemple OLAP AdventureWorks » et nous cliquons sur « Suivant ».

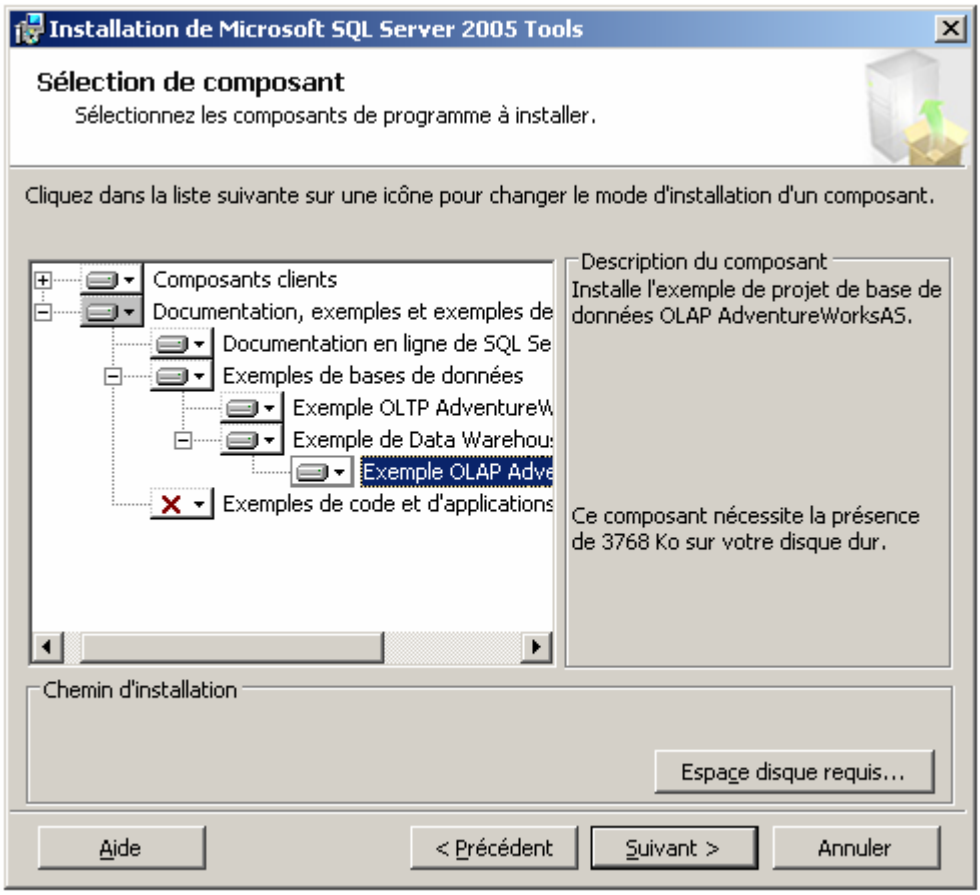


Figure 6-3 : Sélection de composant

A l'écran « Installation des exemples de bases de données » (Figure 6-4), nous cochons l'option « Installer et attacher les exemples de bases de données » et nous sélectionnons sur quelle instance SQL Server nous désirons installer la base de données. Nous validons les choix en cliquant sur Suivant.

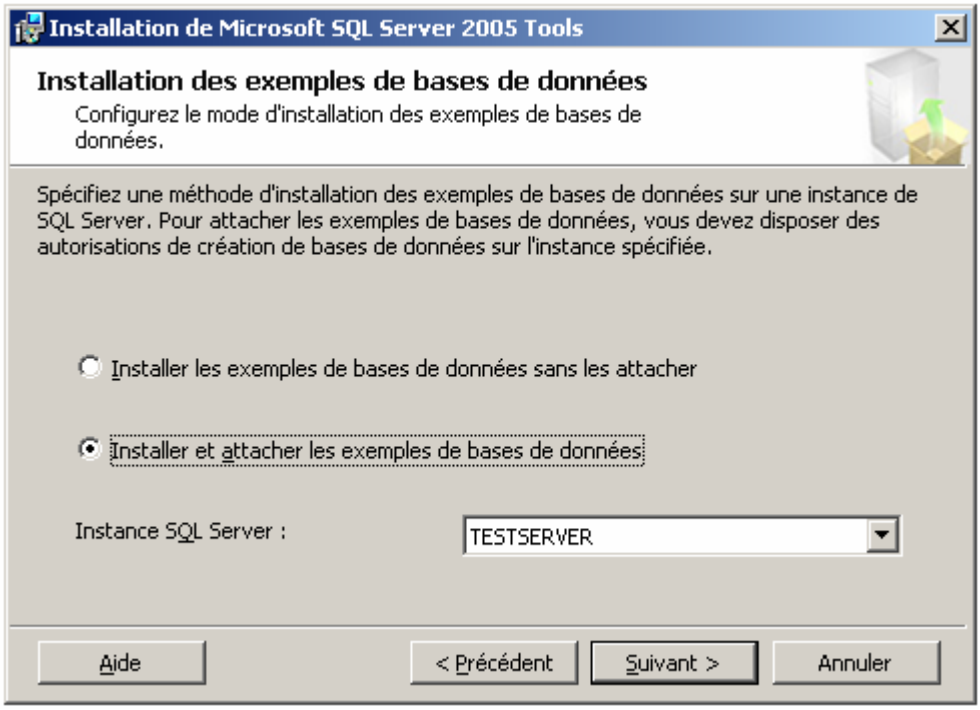


Figure 6-4 : Installation des exemples de bases de données

Arrivé à ce point, l'écran « Prêt pour la mise à jour » (Figure 6-5) nous informe que l'installation est prête à commencer dès que nous cliquons sur « Installer », ce que ne faisons.

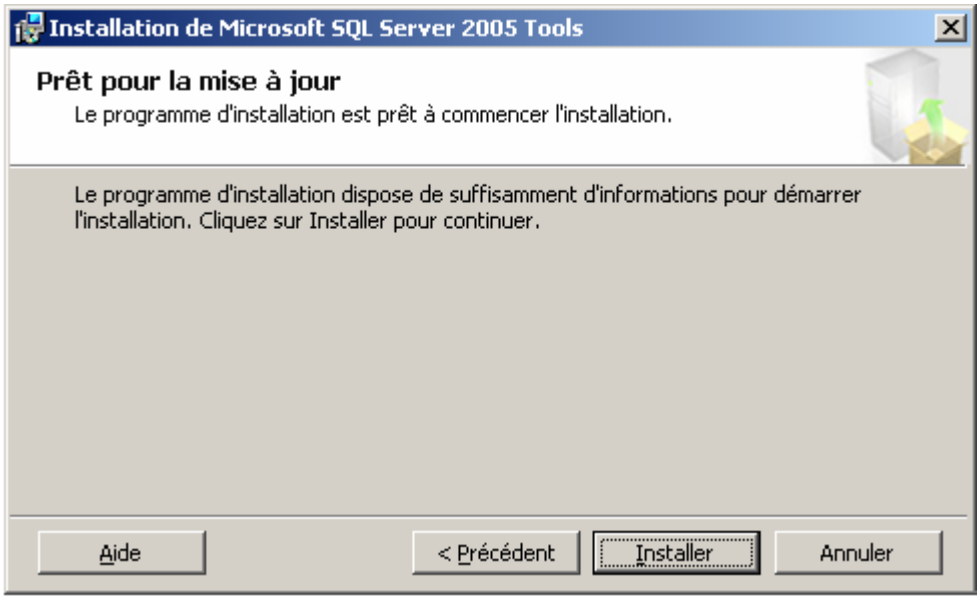


Figure 6-5 : Prêt pour la mise à jour

Une fois l'installation terminée, l'écran « Fin de l'installation de Microsoft SQL Server 2005 » (Figure 6-6) nous en affiche un résumé. Nous cliquons sur « Terminer » pour conclure cette installation

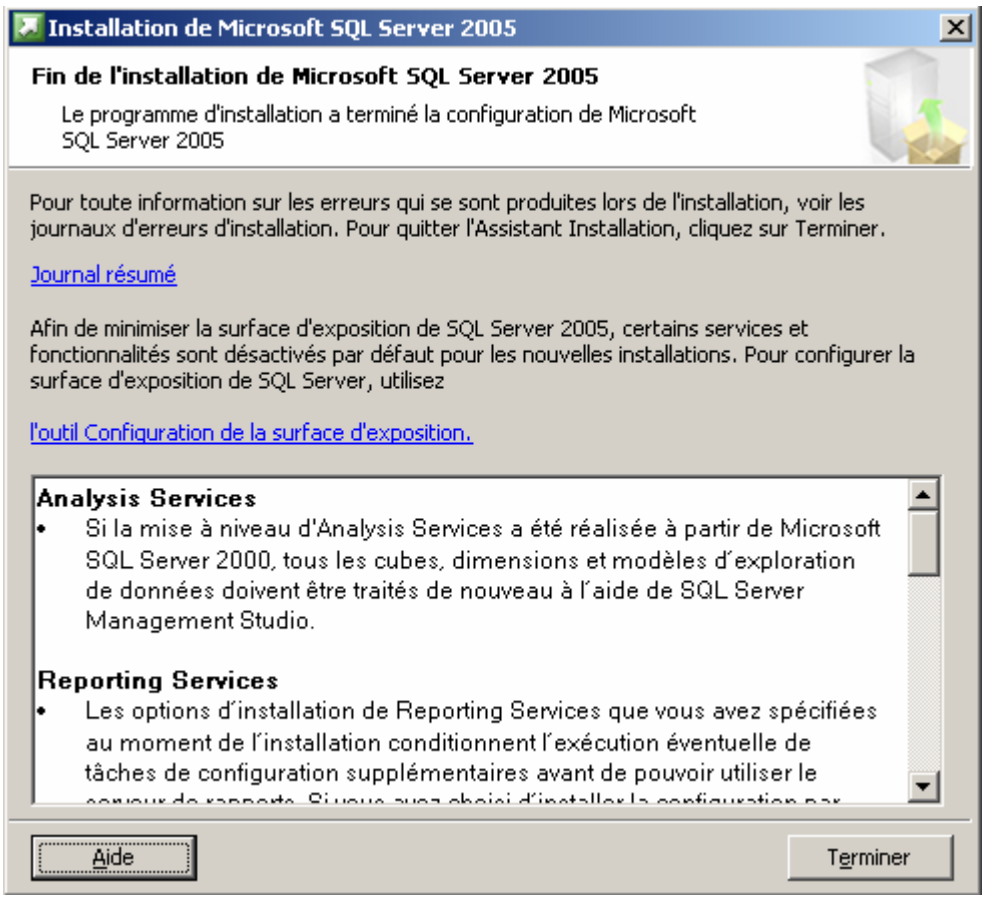


Figure 6-6 : Fin de l'installation de Microsoft SQL Server 2005

6.2 Script SQL

Les scripts SQL suivant doivent être exécutés dans le même ordre que celui figurant dans ce document.

6.2.1 Modification de la table « dbo.ProspectiveBuyer »

6.2.1.1 Ajout du champ « Age »

```
ALTER TABLE [AdventureWorksDW].[dbo].[ProspectiveBuyer] ADD
  Age int NULL
```

Script 6-1

6.2.1.2 Mise à jour du champ « Age »

```
UPDATE [AdventureWorksDW].[dbo].[ProspectiveBuyer]
   SET [Age] = DATEDIFF(YEAR, [BirthDate], '01.01.2007')+1
```

Script 6-2

6.2.1.3 Mise à jour du champ « Education »

```
UPDATE   [AdventureWorksDW].[dbo].[ProspectiveBuyer]
SET      [Education] = 'High School'
WHERE    ([Education]='High Schoo')

UPDATE   [AdventureWorksDW].[dbo].[ProspectiveBuyer]
SET      [Education] = 'Partial College'
WHERE    ([Education]='Partial Co')

UPDATE   [AdventureWorksDW].[dbo].[ProspectiveBuyer]
SET      [Education] = 'Partial High School'
WHERE    ([Education]='Partial Hi')

UPDATE   [AdventureWorksDW].[dbo].[ProspectiveBuyer]
SET      [Education] = 'Graduate Degree'
WHERE    ([Education]='Graduate D')

UPDATE   [AdventureWorksDW].[dbo].[ProspectiveBuyer]
SET      [Education] = 'Bachelors'
WHERE    ([Education]='Bachelors')
```

Script 6-3

6.2.2 Modification de la vue « dbo.vDMPrep »

```
ALTER VIEW [AdventureWorksDW].[dbo].[vDMPrep]
AS
SELECT
  pc.[EnglishProductCategoryName]
  ,Coalesce(p.[ModelName], p.[EnglishProductName]) AS [Model]
  ,c.[CustomerKey]
  ,s.[SalesTerritoryGroup] AS [Region]
  ,DateDiff(YEAR,c.[BirthDate], '01/01/2007')+1 AS [Age]
  ,CASE
     WHEN c.[YearlyIncome] < 40000 THEN 'Low'
     WHEN c.[YearlyIncome] > 60000 THEN 'High'
     ELSE 'Moderate'
   END AS [IncomeGroup]
  ,t.[CalendarYear]
  ,t.[FiscalYear]
  ,t.[MonthNumberOfYear] AS [Month]
  ,f.[SalesOrderNumber] AS [OrderNumber]
  ,f.SalesOrderLineNumber AS LineNumber
  ,f.OrderQuantity AS Quantity
  ,f.ExtendedAmount AS Amount
FROM
  [AdventureWorksDW].[dbo].[FactInternetSales] f
      INNER JOIN [AdventureWorksDW].[dbo].[DimTime] t
        ON f.[OrderDateKey] = t.[TimeKey]
      INNER JOIN [AdventureWorksDW].[dbo].[DimProduct] p
        ON f.[ProductKey] = p.[ProductKey]
      INNER JOIN [AdventureWorksDW].[dbo].[DimProductSubcategory] psc
        ON p.[ProductSubcategoryKey] = psc.[ProductSubcategoryKey]
      INNER JOIN [AdventureWorksDW].[dbo].[DimProductCategory] pc
        ON psc.[ProductCategoryKey] = pc.[ProductCategoryKey]
      INNER JOIN [AdventureWorksDW].[dbo].[DimCustomer] c
        ON f.[CustomerKey] = c.[CustomerKey]
      INNER JOIN [AdventureWorksDW]. [dbo].[DimGeography] g
        ON c.[GeographyKey] = g.[GeographyKey]
      INNER JOIN [AdventureWorksDW].[dbo].[DimSalesTerritory] s
        ON g.[SalesTerritoryKey] = s.[SalesTerritoryKey]
```

Script 6-4

Projet SIMAV

Cahier des charges

Facturation par APDRG : prédiction des recettes des cas non codés

Institut Central des Hôpitaux Valaisans
Av. Grand Champsec 86
Case postale 736
1951 Sion
Suisse

Auteur	: Mathieu Giotta	**Date de création**	: 21.09.2007
Fichier	: PrediRec - 004 Cahier des Charges	**No de version**	: V. finalisée
Etat	:	**Dernière révision**	:
Distribution	:	**Date de distribution**	:
Publication	:		

Table des matières

1 Introduction

1.1 Présentation de la société mandat

Le SIMAV est l'acronyme de Service d'Informatique Médicale et Administrative Valaisan.
Ce service est dirigé par le Réseau Santé Valais (RSV) et doit fournir l'appui informatique nécessaire aux hôpitaux valaisans pour leur bon fonctionnement.

En outre, le SIMAV offre un support dans la réalisation de la plupart des développements des services hospitaliers tel que la comptabilité. De ses nombreuses tâches, nous pouvons citer :

- maintenance réseau ;
- maintenance matérielle ;
- support aux utilisateurs (back office) ;
- développement de petites solutions,

C'est dans cette dernière activité que s'inscrit ce projet.

1.2 Contexte du projet

Depuis 2005, tous les séjours hospitaliers en soins aigus (médecine, chirurgie, gynécologie-obstétrique, pédiatrie, etc.) sont facturés par le RSV sous forme de forfaits liés à la pathologie (APDRG pour All Patients Diagnosis Related Groups). La génération de ces forfaits implique le codage préalable des diagnostics et des interventions documentées dans le dossier médical du patient. Or, lors du bouclement comptable des hôpitaux valaisans, tous les patients sortis durant l'exercice terminé ne sont pas forcements codés, et ne peuvent donc pas être facturés.

Il est nécessaire alors de provisionner les recettes, qui seront perçues après bouclement, afin de les intégrer à l'exercice comptable en cours. Cependant, ces recettes dépendant de la pathologie, elles ne sont pas connues à priori.

1.3 Buts du développement

Le but de ce projet est de déterminer un modèle d'analyse afin d'estimer au mieux les recettes liées aux cas non codés à partir des informations disponibles dans les systèmes opérationnels.
Pour parvenir à provisionner les recettes de ces cas, il a été décidé de mettre en place une solution de Data Mining, que nous appellerons « PrediRec » dans le reste du document.

1.4 Description de l'existant

Actuellement, les hôpitaux calculent leurs provisions pour le secteur somatique aigu sur la base d'un montant moyen forfaitaire de manière empirique afin de prévoir les recettes des cas non codés lors du bouclement comptable. Un entretien avec les comptables aura lieu afin de déterminer précisément leur méthode utilisée actuellement.
Cette solution ne donne pas entière satisfaction, car elle ne tient pas compte de la lourdeur des cas et, de plus, cette méthode est assez fastidieuse.

1.5 Autres aspects généraux

Il faudra tester différents modèles de Data Mining afin d'approximer au mieux les recettes liées au cas non facturés à la fin 2006 et aussi comparer les résultats du Data Mining par rapport à la simulation faite « à la main » pour ces recettes.

2 Intervenant du projet

Henning Mueller	HMU	Professeur responsable
Dr. Alexandre Gnaegi	AGN	Chef du service SIMAV
Thomas Werlen	TWE	Responsable de la simulation
Mathieu Giotta	MGI	Diplômant

3 Demandes fonctionnelles

3.1 Utilisateurs

Les utilisateurs de PrediRec ne sont pas des informaticiens. Il s'agira essentiellement de comptables, de facturistes ainsi que de contrôleurs de gestion. C'est pour cette raison que la solution doit être simple d'utilisation. Ces personnes sont réparties dans les différents sites du RSV : Brigue, Viège, Sierre, Montana, Sion, Martigny, St-Maurice et Monthey.

3.2 Cas d'utilisations

PrediRec pourra être utilisé dans différents contextes :

- lors du bouclement de l'exercice comptable terminé afin de provisionner les recettes des cas non codés
- lors de l'élaboration d'états financiers périodiques, en cours d'exercice, afin de provisionner les recettes des cas non codés
- pour simuler les recettes d'un service ou d'un département connaissant des retards ou des lacunes dans son codage (en raison de vacances, de départs, de sous-dotations, etc.)
- pour proposer la vérification du codage de certains cas, selon un système de controlling

3.3 Fonctionnalités

Fonctionnalités demandées :

- La solution doit permettre à l'utilisateur de choisir les cas pour lesquels la solution doit estimer les recettes parmi les cas présents ou sortis d'un ou plusieurs hôpital (aux).
- L'utilisateur doit pouvoir lui-même exécuter l'analyse des cas.
- La solution doit afficher les recettes totales.
- Les résultats doivent pouvoir être exportés vers Excel.

Fonctionnalités optionnelles :

- Affichage de l'APDRG estimé, de sa valeur en points (cost-weight) et de la valeur du forfait en francs.
- Affichage par rubrique du coût par cas, selon le plan comptable analytique.

3.4 Données sources

Les données sources seront extraites des différents systèmes productifs mis en place actuellement dans les hôpitaux :

- **Le dossier administratif Opale SIAD**
  Dans le dossier administratif se trouvent toutes les données relatives à l'administration du patient : sa date d'entrée et sa date de sortie de l'hôpital, sa durée de séjour, le montant qui est facturé au patient, son APDRG, etc. Lorsqu'il sera codé, c'est également dans ce système que se trouveront les codes diagnostiques et de traitements.
- **Le dossier clinique Phoenix SICL**
  Dans le dossier clinique se trouvent toutes les informations relatives aux traitements du patient : les médicaments prescris durant le séjour, les différentes prestations fournies au patient, ses analyses, etc.
- **Le Data Warehouse DW**
  Le Data Warehouse est alimenté par les 2 systèmes opérationnels ci-dessus (Figure 3-1). Celui-ci est mis à jour quotidiennement, durant la nuit, pour la majorité des données. Certaines données ne sont importées qu'hebdomadairement du fait du travail pour leur chargement.

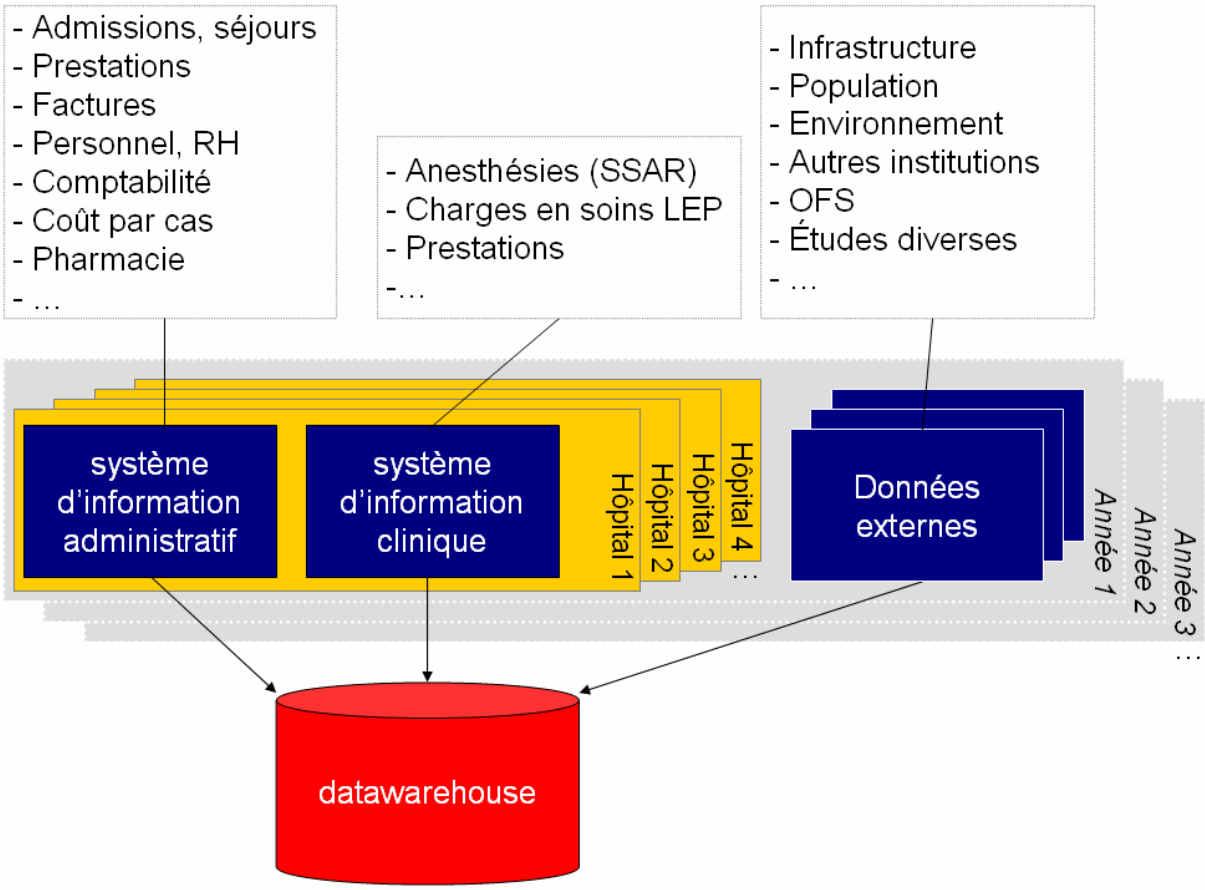


Figure 3-1 :Flux de données du DW

Le schéma suivant représente les flux de données de PrediRec (Figure 3-2).
Les collectes de données se feront principalement à partir du DW et des accès ponctuels pourront être effectués sur les autres sources de données.

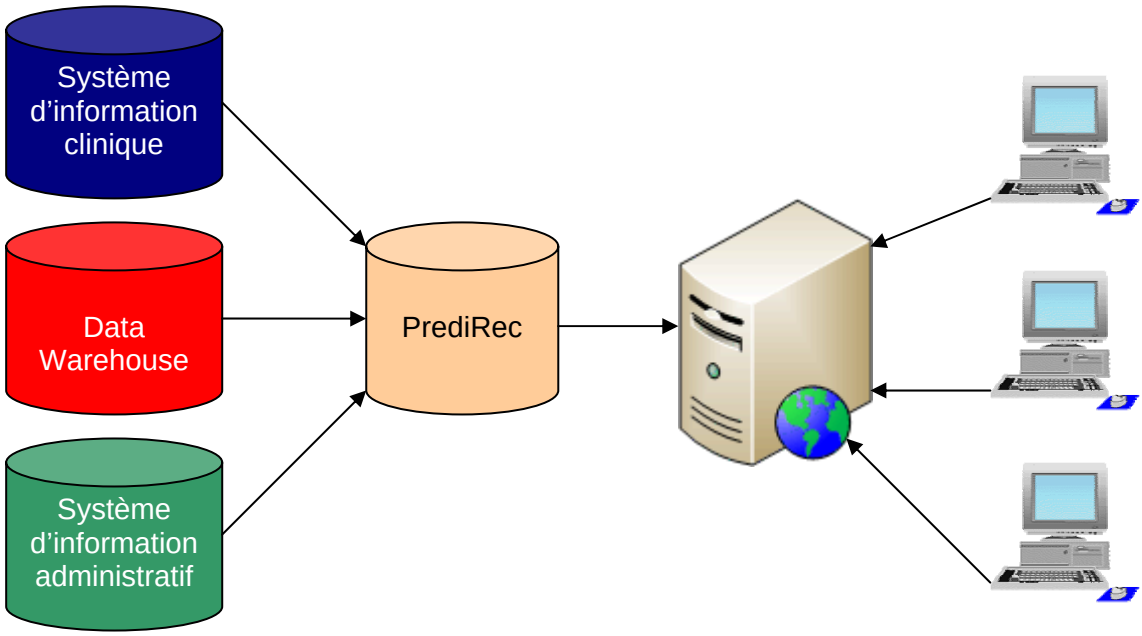


Figure 3-2 : Flux de données de PrediRec

4 Environnement technique

Du fait du nombre de systèmes sources disponibles, les données nécessaires pour l'analyse seront extraites des systèmes productifs afin d'alimenter un Data Mart propre à PrediRec.

L'outil ETL pour alimenter le Data Mart sera celui compris dans la suite SQL Server 2005 c'est-à-dire SQL Server Integration Services (SSIS).
Une attention toute particulière sera portée afin de ne pas perturber les systèmes opérationnels, tant du point de vue fonctionnel que des performances.

Le logiciel de Data Mining pour PrediRec est SQL Serveur 2005 Data Mining de Microsoft, car il est disponible au SIMAV.

Pour les phases de développement, PrediRec sera installé sur une image VMWare Windows Serveur 2003 qui s'exécutera sur un serveur AMD Opterons 2.83 64 bit possédant 16 GB de RAM.
Le Data Mart construit pour les phases sera lui aussi sur l'image VMWare.

Quant aux bases de données sources, elles sont stockées sur différents serveurs et le SGBDR diffère d'une source à l'autre.

- le Data Warehouse
  Serveur : AMD Opterons 2.83 64 bit, 16 GB de RAM
  Base de données : SQL Serveur 2005
- le dossier administratif
  Serveur : Alpha ES80, 2x EV7, 2 GB de RAM
  Base de données : Progress 9
- le dossier clinique
  Serveur : Alpha ES80, 2x EV7, 8 GB de RAM
  Base de données : Oracle 9

5 Assurance Qualité

5.1 Sécurité, accessibilité

5.1.1 Droits d'accès

Seules les personnes habilitées pourront avoir accès à PrediRec.
Le SIMAV est responsable de l'attribution des droits d'accès. Un soin particulier sera porté au respect de la législation et des normes relatives à la protection des données.

MGI assurera la formation des utilisateurs sur l'utilisation du logiciel de Data Mining. MGI sera aussi responsable de l'installation du logiciel sur les postes des utilisateurs finaux et, si nécessaire, installera une solution Internet

Les utilisateurs finaux disposeront d'un accès sécurisé par nom d'utilisateur et mot de passe.

5.1.2 Utilisation des données

Les données provenant des systèmes sources ne doivent pas être modifiées avant leur introduction dans PrediRec.
Les données sources ne doivent pas pouvoir être utilisées à d'autres fins.
Les données introduites dans le logiciel de Data Mining devront être anonymes, c'est-à-dire qu'on ne peut pas retrouver ou deviner l'identité d'un patient à partir de ces données.

5.2 Temps de réponse

Les temps de réponse sont dépendants de la précision des données :

+ de précision => + de données à travailler => + temps de travail CPU

TWE doit définir les temps de réponse maximum souhaités.

5.3 Évolutivité

Actuellement, il n'est prévu aucune évolutivité à court terme. L'introduction généralisée en Suisse d'un nouveau mode de rémunération des hôpitaux, sur la base de forfaits dénommés SwissDRG, est prévue dès 2010.
Par analogie, les hôpitaux, en Valais mais aussi ailleurs en Suisse, pourront reprendre les méthodes de simulation développées dans ce projet.

5.4 Tests

Différents tests de modèles de Data Mining devront être effectués.
Pour chacun des modèles analysés, il faudra comparer avec les recettes simulées avec les recettes réelles des cas non codés.
Pour y arriver, un set de différents cas sera défini, cas pour lesquels nous connaissons déjà les recettes.

Les tests exécutés doivent être documentés afin de faciliter le choix du modèle final. PrediRec devra utiliser le modèle de Data Mining qui permet une approximation au plus proche des recettes réelles.

5.5 Validation

TWE est la personne responsable de la validation du modèle de données, ainsi que du choix des variables.

5.6 Documentation

Une documentation sur l'utilisation de PrediRec doit être fournie à la fin de ce projet, soit :

- le document d'analyse
- le document de design
- les fiches de test
- un manuel d'utilisation
- un rapport de validation par l'utilisateur
- une aide en ligne (optionnel)

De plus, dans le cadre du travail de diplôme, un rapport final et une présentation PowerPoint doivent être livrés.

5.7 Maintenance, mise à jour

Dans un premier temps, la mise à jour des données du Data Mart se fera sur demande des utilisateurs par MGI.

Ensuite, lorsque le système sera mis en place, la mise à jour des données se fera directement par les utilisateurs ou de manière automatique et régulière.

L'utilisateur final est responsable du choix des cas à importer.

6 Plan

Il a été décidé de rendre le rapport final du travail de diplôme pour le 21 décembre 2007
Chaque semaine une feuille de rapport sera saisie et ajoutée au dossier final.

Tâches	Échéance	Personne responsable
Validation du cahier des charges	02/11/2007	HMU, AGN
Validation de l'analyse	13/11/2007	AGN, TWE
Choix du modèle de Data Mining	04/12/2007	AGN, TWE
Fin du travail de diplôme	21/12/2007	MGI

7 Glossaire

7.1 APDRG

Les DRG (Diagnosis Related Groups) sont des systèmes classant les séjours hospitaliers de soins aigus somatiques, sur la base des données récoltées de routine, dans un nombre défini de groupes homogènes du point de vue clinique et du point de vue de la consommation de ressources.

Il existe une multitude de systèmes cousins, dont les plus courants sont les Refined DRGs (RDRG), les All Patient DRGs (APDRG), les All Patient Refined DRGs (APRDRG) ou encore les International-Refined DRGs (IRDRG). La plupart des pays utilisent l'un ou l'autre de ces systèmes, ou ont développé leur propre système national (G-DRG en Allemagne, NordDRG en Scandinavie, ARDRG en Australie, etc.)

Les systèmes DRG ont été introduits aux États-Unis en 1983 pour le financement des soins. Utilisés dans la plupart des pays occidentaux, l'Allemagne et la France ont décidé récemment de leur introduction généralisée. Ces systèmes fournissent un outil performant pour la gestion de l'hôpital, en favorisant la rationalisation des investissements, une meilleure maîtrise des coûts et permettant la comparaison inter-établissements (benchmarking). L'intérêt supplémentaire de ce mode de remboursement est d'être basé sur les données médicales du patient et donc de tenir compte du coût du traitement, contrairement au forfait par jour traditionnel.

Introduits et testés en Suisse en 1998, les APDRG sont utilisés par un nombre croissant d'hôpitaux. Il est prévu qu'en 2006 près d'un hôpital sur deux facturera ses prestations sur cette base. Le projet de recherche et développement APDRG 1998-2004 avait pour but d'adapter cette technique aux besoins suisses et de réaliser sa mise en application. Les premières utilisations dans quelques cantons ces dernières années ont plus que confirmé les espoirs des initiateurs. Le nouveau club prend la relève de ce projet, pour en assurer la maintenance régulière et encourager la coopération entre ces utilisateurs.

Le succès du projet APDRG a incité les autorités et partenaires sanitaires de lancer un nouveau projet de recherche et de développement, SwissDRG 2004-2007. Ce projet a pour ambition de poursuivre la réflexion sur le financement des hôpitaux et d'identifier les solutions les plus judicieuses pour la Suisse dans le futur.
Source : www.apdrgsuisse.ch

7.2 Le codage

Transcription des diagnostics et des interventions décrits dans le dossier médical du patient, essentiellement dans la lettre de sortie et le rapport opératoire. Les diagnostics sont codés en Suisse selon la Classification statistique internationale des maladies et des problèmes de santé connexes, 10ème révision (CIM-10), publiée par l'OMS, alors que les interventions et les traitements sont codés selon la Classification suisse des interventions chirurgicale (CHOP), qui est une adaptation de la classification américaine ICD-9-CM, Volume 3. En Suisse, une nouvelle version de la CHOP est publiée chaque année par l'Office fédéral de la statistique. En 2007, c'est la version 9.0 qui a cours.

L'obligation du codage faite aux hôpitaux et l'utilisation des classifications de référence figurent dans une annexe de l'Ordonnance fédérale du 30 juin 1993 concernant l'exécution des relevés statistiques fédéraux, en particulier la statistique médicale des hôpitaux. Cette ordonnance accompagne la Loi fédérale du 9 octobre 1992 sur la statistique fédérale, qui met sur pied et motive les différents relevés statistiques dans le domaine de la santé.

La statistique médicale des hôpitaux, qui impose le codage dans tous les hôpitaux suisses, avait quatre buts principaux:
Surveillance épidémiologique (incidence et prévalence des maladies, état de santé de la population et mesures préventives ou thérapeutiques)
Saisie de prestations médicales homogènes et contrôle de la qualité
Bases pour la planification intra- et inter-cantonale
Mise à disposition de données pour la recherche et publications
Avec l'introduction de la Loi fédérale sur l'assurance-maladie (LaMal) du 18.03.1994 et son Ordonnance sur le calcul des coûts et le classement des prestations par les hôpitaux et les établissements médico-sociaux dans l'assurance-maladie (OCP) de 2003, l'accent a surtout été mis sur le deuxième objectif. Ainsi, le codage médical a été la base du financement des hôpitaux par pathologie, selon le système APDRG (All Patient Diagnosis Related Groups), dès 2002 dans le canton de Vaud et dès 2004/2005 en Valais et dans d'autres cantons. La révision en cours de la LaMal prévoit d'ailleurs la généraliation d'un tel mode de financement dans tous les hôpitaux (SwissDRG) dès 2010/2011.

7.3 Cost-Weight (CW)

Chaque groupe de pathologie a son propre poids appelé cost weight (CW). Le cost weight indique le poids des frais de traitement moyens des patients d'un groupe DRG par rapport à celui de l'ensemble des patients en traitement stationnaire aigu en Suisse. Il est par exemple de 3,067 pour le groupe APDRG 191 (shunt intra-abdominal et interventions sur le pancréas et le foie, avec cc [comorbidités et/ou complications]) et de 1,432 pour le groupe APDRG 554 (interventions sur hernie, avec cc majeure). En d'autres termes, on suppose que le traitement d'un patient du groupe 191 coûte en moyenne 2,14 fois plus que celui d'un patient du groupe 554 et 3,067 fois plus que le traitement d'un patient moyen en traitement stationnaire aigu en Suisse. De la même façon, on part du principe que les frais de traitement d'une intervention vasculaire extracranienne (APDRG 5, CW 1,339) sont environ 2,6 fois supérieurs à ceux d'une stupeur et d'un coma traumatiques de moins d'une heure chez un patient âgé de moins de 18 ans (APDRG 763, CW 0,515).
Source :
http://www.zmt.ch/fr/stationaere_tarife/stationaere_tarife_apdrg/stationaere_tarife_apdrg_grundlageninformationen.htm

7.4 Data Ming

De manière générale, on peut le définir comme l'extraction d'informations ou de connaissances originales, auparavant inconnues, potentiellement utiles à partir de gros volumes de données (d'après Frawley et Piateski-Shapiro).

Selon SAS-INSTITUTE, il s'agit du processus de sélection, exploration, modification et modélisation de grandes bases de données afin de découvrir des relations entre les données jusqu'alors inconnues.

Le Data Mining correspond donc à l'ensemble des techniques et des méthodes qui à partir de données permettent d'obtenir des connaissances exploitables. Son utilité est grande dès lors que l'entreprise possède un grand nombre d'informations stockées sous forme de bases de données.

Plus particulièrement, une distinction plus précise s'établit autour du concept de KDD (Knowledge Discovery in Database ou Découverte de Connaissances dans les Bases de Données) et celui de Data Mining. En effet, ce dernier n'est que l'une des étapes du processus de découverte de connaissances correspondant précisément à l'extraction des connaissances à partir des données. Avant de réaliser une étude Data Mining, il faut donc procéder à l'élaboration d'un Data Warehouse (Entrepôt de Données).

En outre, bien qu'utilisant des techniques et une démarche statistique, le Data Mining et ses outils sont appelés à être utilisés par des non-statisticiens praticiens spécialistes du problème à modéliser. Pour cela, le progiciel utilisé doit avoir des caractéristiques spécifiques (Cf question 9).

Les applications du Data Mining sont multiples, elles concernent: la grande distribution, la vente par correspondance, les opérateurs de télécommunications, les banques et assurances, etc. Le domaine majeur où le Data Mining a prouvé son efficacité est la gestion de la relation client (CRM ou Customer Relationship Management). En effet, le Data Mining permet par une meilleure connaissance de la clientèle d'accroître les ventes.
Source : http://www.web-datamining.net/forum/faq.asp

Projet SIMAV

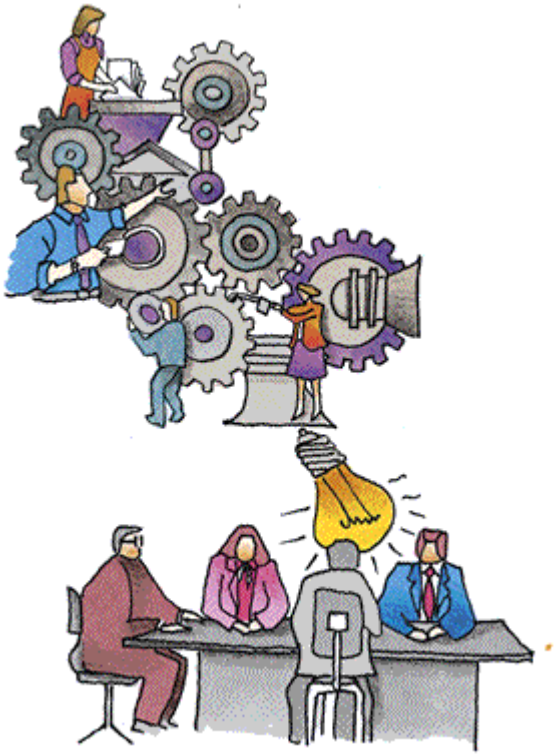
Document d'analyse

Facturation par APDRG : prédiction des recettes des cas non codés

Institut Central des Hôpitaux Valaisans
Av. Grand Champsec 86
Case postale 736
1951 Sion
Suisse

Auteur	: Mathieu Giotta	**Date de création**	: 13.11.2007
Fichier	: PrediRec - 005 Document d'analyse	**No de version**	: V. finalisée
Etat	:	**Dernière révision**	:
Distribution	:	**Date de distribution**	:
Publication	:		

Table des matières



1 Introduction

1.1 Présentation générale

SIMAV est l'acronyme du Service d'Informatique Médicale et Administrative Valaisan.
Ce service est dirigé par le Réseau Santé Valais (RSV) et doit fournir l'appui informatique nécessaire aux hôpitaux publics valaisans pour leur bon fonctionnement.

En outre, le SIMAV offre un support dans la réalisation de la plupart des développements des services hospitaliers, dont la comptabilité par exemple. De ses nombreuses tâches, nous pouvons citer :

- maintenance réseau
- maintenance matérielle
- support aux utilisateurs (back office)
- développement de petites solutions

C'est dans cette dernière activité que s'inscrit ce projet.

Depuis 2005, les séjours hospitaliers en soins somatiques aigus sont facturés par le RSV sous forme de forfaits liés à la pathologie du patient (APDRG). Pour que ces forfait soient générés et que les séjours puissent être facturé, les diagnostics et les traitements indiqués dans le dossier médical doivent être codés par du personnel spécialisé. Cependant, lors des bouclements comptables, les dossiers des patients sortis durant l'exercice écoulé ne sont pas forcément tous codés et, par extension, pas facturés.
Or, ces cas doivent être provisionnés afin que le bouclement comptable soit le plus correct possible.

Actuellement, la méthode utilisée pour estimer ces cas est une méthode empirique et contraignante. De plus, cette méthode ne tient pas compte de la lourdeur des cas, ce qui fait qu'au final, les provisions estimées ne sont pas toujours au plus proche de la réalité.

Ce projet a comme but de mettre en place un modèle d'analyse afin d'approximer au mieux les recettes des cas non codés à l'aide d'un l'outil de Data Mining. Un outil de Data Mining permet, à partir d'un set de donnée d'extraire des relations entre les données ou d'effectuer un modèle d'analyse de prédiction.

L'application développée se nommera PrediRec et sera installée sur les mêmes serveurs que ceux du Data Warehouse du SIMAV.

1.2 Acronymes, abréviation et définition

SIMAV : Service d'Informatique Médicale et Administrative Valaisan

RSV : Réseau Santé Valais

CW : Cost-Weight
Chaque groupe de pathologie a son propre poids relatif, exprimé en points et appelé cost-weight (CW). Le cost-weight indique le rapport entre les coûts moyens de traitement des patients d'un DRG donné et les coûts moyens de traitement de l'ensemble des patients en traitement stationnaire aigu en Suisse. Corollairement, plus le cost-weight d'un DRG est élevé, plus les coûts de prise en charge des patients classés dans ce DRG sont élevés. Le cost-weight est donc un indicateur de la lourdeur relative d'un séjour hospitalier.

CW pondéré : Dans chaque DRG, les patients ont des durées de séjour variables, distribuées statistiquement selon une courbe normale (ou de Gauss). Les séjours dépassant un certain percentile (méthode Gamma), variable pour chaque DRG, sont appelés « cas extrêmes » ou outliers. Un outlier est dit supérieur si sa durée de séjour est anormalement élevée et inférieur si elle est anormalement courte. Chaque jour dépassant la limite théorique (la borne ou trim point) est pondéré positivement ou négativement en plus du cost-weight brut. Le résultat de cette pondération est le cost-weight pondéré. Le cost-weight pondéré d'un outlier inférieur sera donc plus petit que le cost-weight brut de ce DRG, alors que c'est l'inverse pour un outlier supérieur. Pour tous les cas situés à l'intérieur des trim points (environ 90% du total), les deux valeurs seront identiques. C'est le cost-weight pondéré qui est utilisé pour le calcul du forfait facturé.

PrediRec : nom du projet, concaténation des termes Prediction et Recette

APDRG : All Patient DRGs. C'est l'un des nombreux systèmes de DRG.
Les DRG (Diagnosis Related Groups) sont des systèmes classant les séjours hospitaliers de soins aigus somatiques, sur la base des données récoltées de routine, dans un nombre défini de groupes homogènes du point de vue clinique et du point de vue de la consommation de ressources.

TARMED : TARMED est le nouveau tarif des prestations, valable pour toutes les prestations médicales ambulatoires à l'hôpital et dans le cabinet médical. Ce projet ambitieux a commencé avec la révision totale des tarifs médicaux (GRAT). Avec l'admission de la loi sur l'assurance-maladie par le peuple Suisse en 1994 il a pris une toute autre dimension. L'art.43, alinéa 5 de LAMal prévoit que les tarifs des prestations médicales dans le domaine de l'assurance maladie se basent sur uns structure de tarifs conclue pour toute la Suisse. La révision totale des tarifs des prestations hospitalières a commencé en 1997. De la fusion de ces deux projets est né TARMED en 1999.
En 2002 le conseil fédéral a approuvé la structure des tarifs TARMED 1.1r. Les assureurs accidents, militaire et invalidité ont introduit le tarif dès le 1er mai 2003. Depuis le 1er janvier 2004 le TARMED est appliqué globalement.

Attribut discret : Ce dit d'une variable qui possède un état fini de valeur.
Ex. Sexe : masculin ou féminin.

Attribut continu : Ce dit d'une variable qui représente un jeu contenu de données numériques.
Ex. le revenu mensuel, le total d'une facture.

Modèle d'exploration de données, modèle d'analyse
: L'exploration de données est fréquemment définie comme « le processus d'extraction d'informations valides, authentiques et utilisables à partir de bases de données de grande taille ». En d'autres termes, l'exploration de données dégage les modèles et les tendances existant dans les données. Ces modèles et tendances peuvent être collectés ensemble et définis en tant que modèle d'exploration de données.

Perceptron : Le *perceptron* est un modèle de réseau de neurones avec algorithme d'apprentissage créé par Frank Rosenblatt en 1958.

1.3 Documents de référence

SIMAV	:	http://www.ichv.ch/default.asp?contentID=662
RSV	:	http://www.rsv-gnw.ch/
APDRG	:	www.apdrgsuisse.ch
CW	:	http://www.zmt.ch/fr/stationaere_tarife/stationaere_tarife_apdrg/stationaere_tarife_apdrg_grundlageninformationen.htm
TARMED	:	http://www.tarmedsuisse.ch/
SSAS	:	http://technet.microsoft.com
TechNet	:	http://technet.microsoft.com
Perceptron	:	http://www.grappa.univ-lille3.fr/~gilleron/PolyApp/node19.html

2 Description générale de PrediRec

L'outil doit permettre à un nombre limité d'utilisateur tel que des comptables ou des facturistes de sélectionner et d'estimer des cas non codés. Pour le faire, l'utilisateur doit pouvoir choisir quelle la variable qu'il désire évaluer.
PrediRec doit aussi offrir la possibilité à chacun des utilisateurs de mettre à jour ses modèles de prévisions à l'aide des données les plus récentes dans les systèmes sources.
Les utilisateurs doivent pourvoir consulter leurs résultats directement à l'écran, mais aussi les exporter dans MS Excel.

Comme l'essentiel de ce travail de diplôme consiste à apprendre et maîtriser les concepts de base du Data Mining et de son application/intégration en entreprise, les fonctionnalités et l'ergonomie de l'interface utilisateur sont réduites à un minimum, mais suffisantes pour que l'application PrediRec soit utilisable.

2.1 L'outil de Data Mining

Le logiciel de Data Mining choisi est SQL Serveur Analysis Services (SSAS) car il est disponible au SIMAV.

Cet outil comporte plusieurs algorithmes, listés dans le Tableau 2-1, pouvant être utiles pour ce projet.

Nom		
Microsoft Decision Trees (MDT)	Utilisation	algorithme de prédiction d'attribut discret algorithme de prédiction d'attribut continu recherche de groupe d'éléments communs
	Définition (TechNet)	L'arbre de décision est un algorithme de classification et de régression utilisé pour la modélisation prédictive d'attributs discrets et continus. Pour la prédiction d'attributs discrets, il effectue des prévisions en fonction des relations entre les colonnes d'entrée. Pour la prédiction d'attributs continus, il utilise la régression linéaire.
Microsoft Naive Bayes (MNB)	Utilisation	algorithme de prédiction d'attribut discret
	Définition (TechNet)	L'algorithme MNB est un algorithme de classification qui est conçu pour la modélisation prédictive. Cet algorithme calcule la probabilité conditionnelle entre les colonnes d'entrée et les colonnes prévisibles, et suppose que les colonnes sont indépendantes. C'est en raison de cette supposition d'indépendance que l'algorithme s'appelle algorithme bayésien naïf (Naive Bayes). En effet, la supposition est souvent naïve étant donné que, en faisant cette supposition, l'algorithme ne prend pas en compte les dépendances qui peuvent exister.

Clusters Microsoft	Utilisation	algorithme de prédiction d'attribut discret recherche de groupe d'éléments similaires
	Définition (TechNet)	L'algorithme Clusters Microsoft est un algorithme de segmentation. L'algorithme utilise des techniques itératives pour grouper les cas d'un jeu de données en clusters contenant des caractéristiques similaires. Ces groupements sont utiles pour l'exploration des données, l'identification d'anomalies dans les données et la création de prévisions.
Microsoft Neural Network (MNN)	Utilisation	algorithme de prédiction d'attribut discret
	Définition (TechNet)	L'algorithme MNN crée des modèles d'exploration de données de classification et de régression en construisant un réseau perceptron multicouche de neurones. Similaire à l'algorithme MDT, l'algorithme calcule les probabilités de chaque état possible de l'attribut d'entrée lorsque chaque état de l'attribut prévisible lui est fourni. Vous pouvez par la suite utiliser ces probabilités pour prédire le résultat de l'attribut prédit en fonction des attributs d'entrée.
Microsoft Time Series (MTS)	Utilisation	algorithme de prédiction continu
	Définition (TechNet)	L'algorithme MTS est un algorithme de régression qui est conçu pour la création de modèles d'exploration de données permettant de prédire des colonnes continues, telles que des ventes de produits, dans un scénario de prévision. Alors que d'autres algorithmes Microsoft créent des modèles, tels que les modèles d'arbre de décision, qui se basent sur des colonnes d'entrée pour prédire la colonne prévisible, la prédiction dans un modèle de série chronologique est uniquement basée sur les tendances que l'algorithme dégage dans les données d'origine pendant la création du modèle.
Microsoft Sequence Clustering (MSC)	Utilisation	prévision d'une séquence recherche de groupe d'éléments similaires
	Définition (TechNet)	L'algorithme MSC est un algorithme d'analyse de séquence. Cet algorithme vous permet d'explorer des données qui contiennent des événements qui peuvent être liés en suivant des chemins ou des séquences. L'algorithme recherche les séquences les plus communes en groupant, ou en regroupant en clusters, les séquences identiques. Ces séquences peuvent être d'une forme quelconque.

Algorithme Microsoft Association	Utilisation	recherche d'éléments communs
	Définition (TechNet)	L'algorithme Microsoft Association est un algorithme d'association qui est utile pour les moteurs de recommandation. Un moteur de recommandation recommande des produits aux clients en se basant sur les éléments qu'ils ont déjà achetés ou pour lesquels ils ont manifesté un intérêt. L'algorithme Microsoft Association est utile également pour l'analyse d'un panier d'achats.

Tableau 2-1 : Liste des algorithmes présent dans SSAS

Une première analyse nous a permis de pré-choisir les algorithmes servant à la prédiction d'attribut discret ou continu :

- Microsoft Decision Trees ;
- Microsoft Naive Bayes ;
- Clusters Microsoft ;
- Microsoft Neural Network.

L'algorithme « Microsoft Time Series » n'est pas retenu car il ne correspond pas aux attentes de ce projet.

2.2 Vue d'ensemble des fonctionnalités

PrediRec sera en mesure de proposer, au minimum, les fonctionnalités suivantes :

Choix des cas non codés
L'utilisateur final doit pouvoir lui-même sélectionner les cas non codés qu'il désire estimer. PrediRec doit être en mesure de se connecter aux bases de données sources afin d'y extraire les informations relatives aux cas non codés.

Estimation des cas non codés
L'utilisateur doit avoir la possibilité de choisir, afin d'estimer ses cas non codés, des modèles d'analyse correspondant à différentes années.

Mise à jour des modèles d'exploration de données
La mise à jour des modèles d'analyse sera réalisée par l'utilisateur de PrediRec.
Chaque utilisateur possèdera son propre modèle, c'est-à-dire qu'un utilisateur x peut mettre à jour son modèle de données pendant que l'utilisateur y peut travailler avec le sien. De cette manière, un utilisateur peut évaluer ses cas non codés sans risquer d'être déstabilisé par une mise à jour impromptue du modèle.

Exportation des cas simulés dans MS Excel
A la suite d'une simulation de cas non codés par l'utilisateur, celui-ci doit pouvoir exporter ses résultats dans MS Excel.

Simulation préventive d'un cas isolé
L'utilisateur de PrediRec aura la possibilité de simuler un cas dans le but, par exemple, d'effectuer un devis avant une hospitalisation.

2.3 Requis non fonctionnels

Chacun des modèles d'analyse sera lié à un nom d'utilisateur ainsi qu'à l'année correspondante au modèle d'analyse.

2.4 Hypothèse

Le système source principal sera le Data Warehouse. Si une donnée nécessaire n'est pas disponible dans celui-ci, les personnes en charge du DW seront habilitées à reprendre

l'information depuis les autres systèmes sources afin de la mettre à disposition des utilisateurs de PrediRec.

3 Description détaillée

3.1 Use Case

3.1.1 Diagramme de contexte

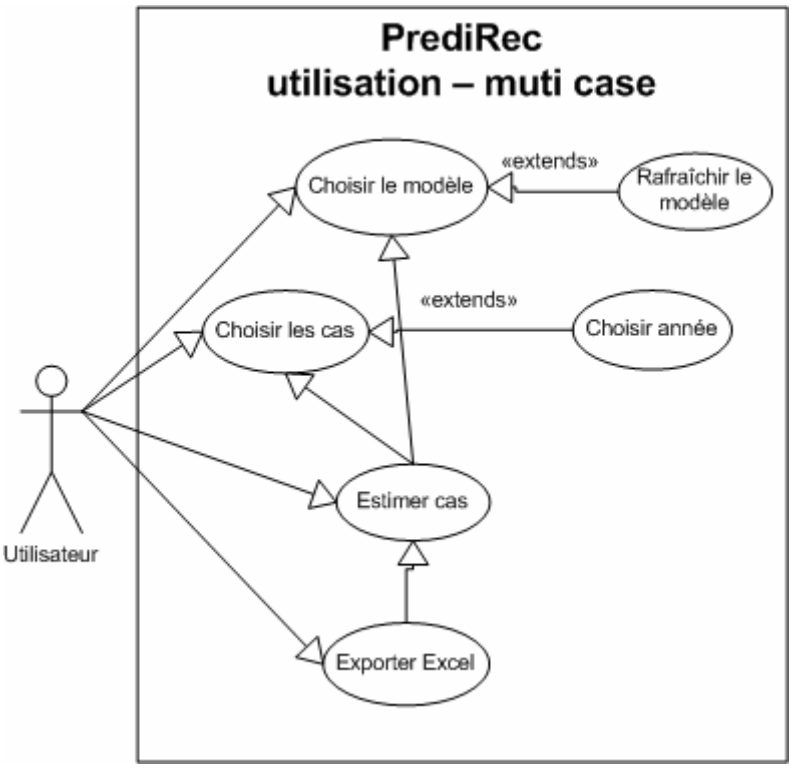


Figure 3-1 : Use case – Multi-Case

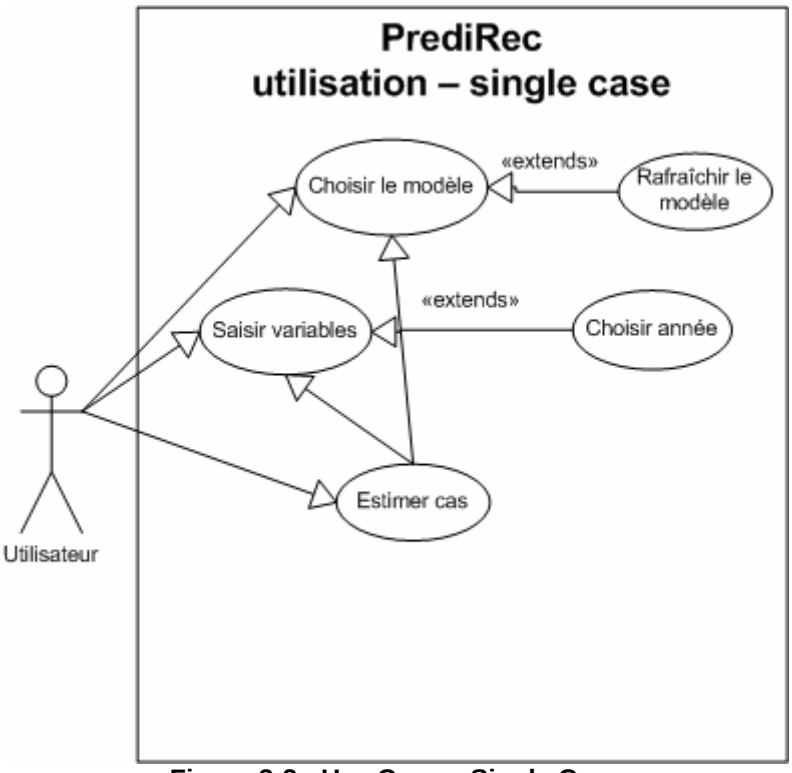


Figure 3-2 : Use Case – Single Case

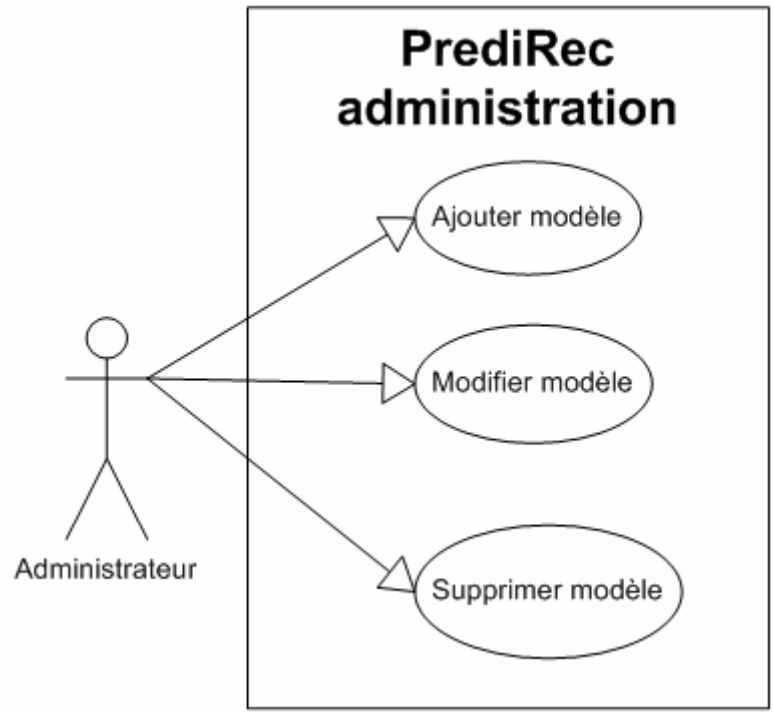


Figure 3-3 : Use Case - Administration

3.1.2 Liste des Use Case

PrediRec – Utilisation Multi-Case

Dans ce scénario, l'utilisateur peut choisir le modèle d'exploration de données, les cas sortis et non codés ainsi que les cas non sortis.

Une fois les cas choisis, l'utilisateur peut les estimer et, si besoin, les exporter dans MS Excel.

PrediRec – Utilisation Single-Case
Dans ce scénario, l'utilisateur peut créer un cas fictif en renseignant chaque variable nécessaire à cette simulation et le soumettre au modèle d'exploration de données choisi.

PrediRec – Administration
Dans ce scénario, l'administrateur peut ajouter, modifier ou supprimer des modèles d'exploration de données.

3.1.3 Description du Use Case : PrediRec – Utilisation Multi Case

Etant donné que la principale fonctionnalité du côté utilisateur est la prédiction des recettes liées aux cas non codés, voici une description plus précise de ce scénario.

L'utilisateur, une fois connecté sur le système PrediRec, a la possibilité de choisir la variable à prédire : soit le CW, soit le total facturé ou le CW pondéré.
Ensuite, il doit sélectionner l'année de référence du modèle, car un modèle de CW ou de CW pondéré peut être utilisé d'une année à l'autre tandis que le total facturé non. Le total facturé est égal au CW pondéré multiplié par la valeur du point du CW qui est réévaluée chaque année.

De plus, l'utilisateur doit choisir s'il désire :

- les cas non codés et sortis de l'hôpital en choisissant l'année de sortie dans une liste déroulante
- les cas pas encore sorti de l'hôpital via une case à cocher

Nb. L'utilisateur peut cumuler les deux options ci-dessus.

- optionnel :
  L'utilisateur peut charger une liste de numéros de patients depuis une feuille MS Excel.

Une fois ces choix effectués, PrediRec effectue une requête les bases de données source et en extrait une liste de cas qu'il présente sous forme de tableau.
A partir de cette liste, l'utilisateur a la possibilité de choisir lesquels des cas présentés il désire estimer. Par défaut, tous les cas sont sélectionnés.
Une fois les cas choisis, l'utilisateur peut lancer la simulation.
A la fin de celle-ci, le système affiche pour chaque cas la valeur estimée et offre la possibilité de les exporter vers MS Excel.

3.2 Diagramme de séquence

Ci-dessous (Figure 3-4), le diagramme de séquence de l'Use-Case PrediRec - MultiValue

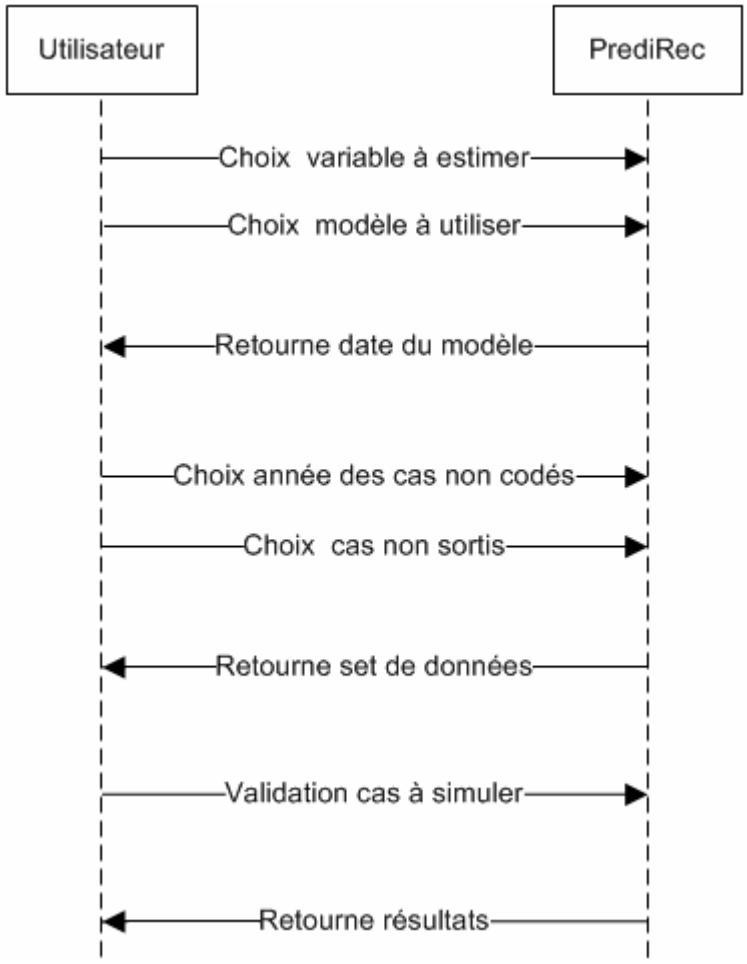


Figure 3-4 : Diagramme de séquence

3.3 Sources de données

PrediRec puise les variables nécessaires à la création du modèle d'analyse dans le Data Warehouse, car celles-ci y sont nettoyées et corrigées.
Le Data Warehouse contient des données provenant des systèmes clinique et administratif des hôpitaux valaisans :

- **Le dossier administratif Opale SIAD**
  Dans le dossier administratif se trouvent toutes les données relatives à l'administration du patient : sa date d'entrée et sa date de sortie de l'hôpital, sa durée de séjour, le montant qui est facturé au patient, son APDRG, etc. Lorsqu'il sera codé, c'est également dans ce système que se trouveront les codes diagnostiques et de traitements.
- **Le dossier clinique Phoenix SICL**
  Dans le dossier clinique se trouvent toutes les informations relatives aux traitements du patient : les médicaments prescris durant le séjour, les différentes prestations fournies au patient, ses analyses, etc.

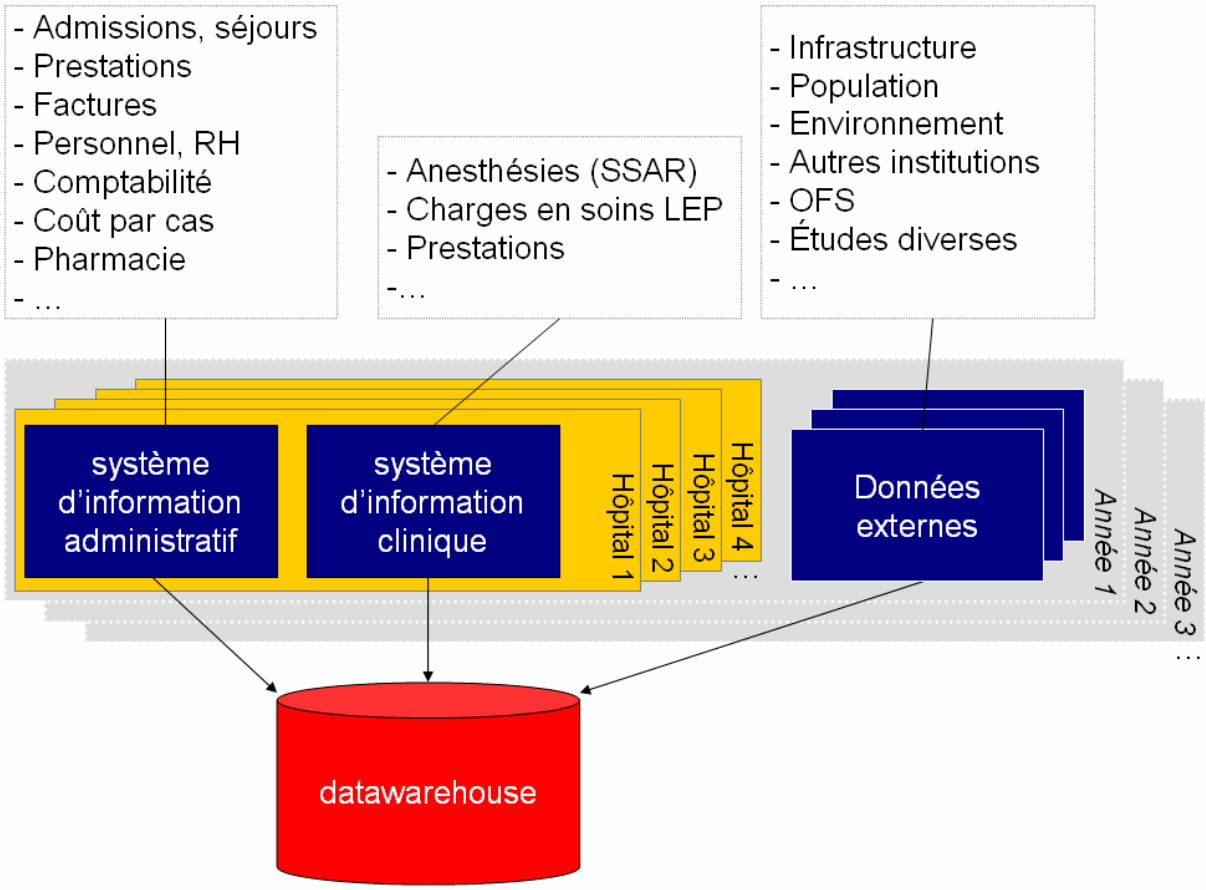


Figure 3-5 : Flux de données du DW

Dans l'éventualité où une variable n'est pas disponible dans le Data Warehouse, l'équipe en charge de celui-ci est habilitée à aller chercher dans les systèmes SIAD et SICL l'information nécessaire.

3.4 Variables potentiellement utilisables

Les variables pré choisies pour la mise en place des modèles d'analyse sont décrites dans le Tableau 3-1.
Le choix de ces variables a été effectué durant les divers entretiens avec les codificatrices et les facturistes.

Nom de la variable	Connu uniquement à la sortie du patient	Remarques
Type d'admission		7 codes différents
Mode d'entrée		24 codes différents
Provenance		84 codes différents
Décision d'envoi		8 codes différents
Genre d'admission		9 codes différents
Mode de sortie	X	35 codes différents
Destination	X	87 codes différents
Prise en charge après la sortie	X	9 codes différents
Sexe		2 codes différents
Résidence pour convention		5 codes différents
Age à l'entrée		Calculé
Cas		49 codes différents
Classe		8 codes différents

Nom de la variable	Connu uniquement à la sortie du patient	Remarques
Type de patient		31 codes différents
Tarif		44 codes différents
Type de taxe		39 codes différents
Groupe de classe		3 codes différents
Division		35 codes différents
Service		114 codes différents
Unité		77 codes différents
Médecin traitant		169 codes différents
Spécialité du médecin traitant		29 codes différents
Genre de médecin traitant		12 codes différents
Durée de séjour		Calculé
Durée de séjour nette		Calculé
Liste des prestations Tarmed		4770 codes différents
Points de prestation Tarmed		Calculé
Valeur des prestations Tarmed		Calculé
CW		Calculé
CW pondéré		Calculé
Total facturé		Calculé
Heures de soins intensifs		Calculé
Lettre de sortie	X	Texte libre
Diagnostic principal		Texte libre
Diagnostic secondaire		Texte libre
Comorbidités		Texte libre
Intervention principale		Texte libre
Intervention secondaire		Texte libre

Tableau 3-1 : Liste des variables potentielles

Toutes ces variables ne doivent pas forcément être utilisées dans les modèles d'analyse mais peuvent être sélectionnées au moment de la création d'un modèle d'exploration de données.

3.5 Variables à estimer

Les variables que PrediRec doit estimer sont les suivantes :

Nom de la variable	Type d'attribut	Remarques
Total Facturé	continu	
CW	discret	optionnel
CW pondéré	continu	

Tableau 3-2 : Liste des variables à estimer

Au maximum, il y autant de CW différents qu'il y a d'APDRG (environ 700). Par contre, il y a potentiellement une infinité de CW pondérés différents, car ceux-ci dépendent des durées de séjour.

3.6 Méthode d'authentification

L'authentification n'est pas à la charge de PrediRec car il prévu que cette application soit déposée sur un site Intranet possédant déjà un système d'authentification par nom d'utilisateur et mot de passe.

3.7 Installation et déploiement

Les pages ASP auxquelles les utilisateurs ont accès doivent être mis en place dans l'Intranet du Data Warehouse, c'est à dire Business Objects InfoView XI R2, communément appelé WEBI. Il s'agit d'une application WEB fonctionnant en https sur un serveur IIS 5.

Le moteur de Data Mining, SSAS, doit être, quant à lui, installé sur un des serveurs de base de données du Data Warehouse.

4 Interfaces utilisateurs

Les textes de l'interface utilisateur sont bilingues : français et allemand.

4.1 Interface Multi-Case

Exemple de l'interface pour le choix de cas multiple :

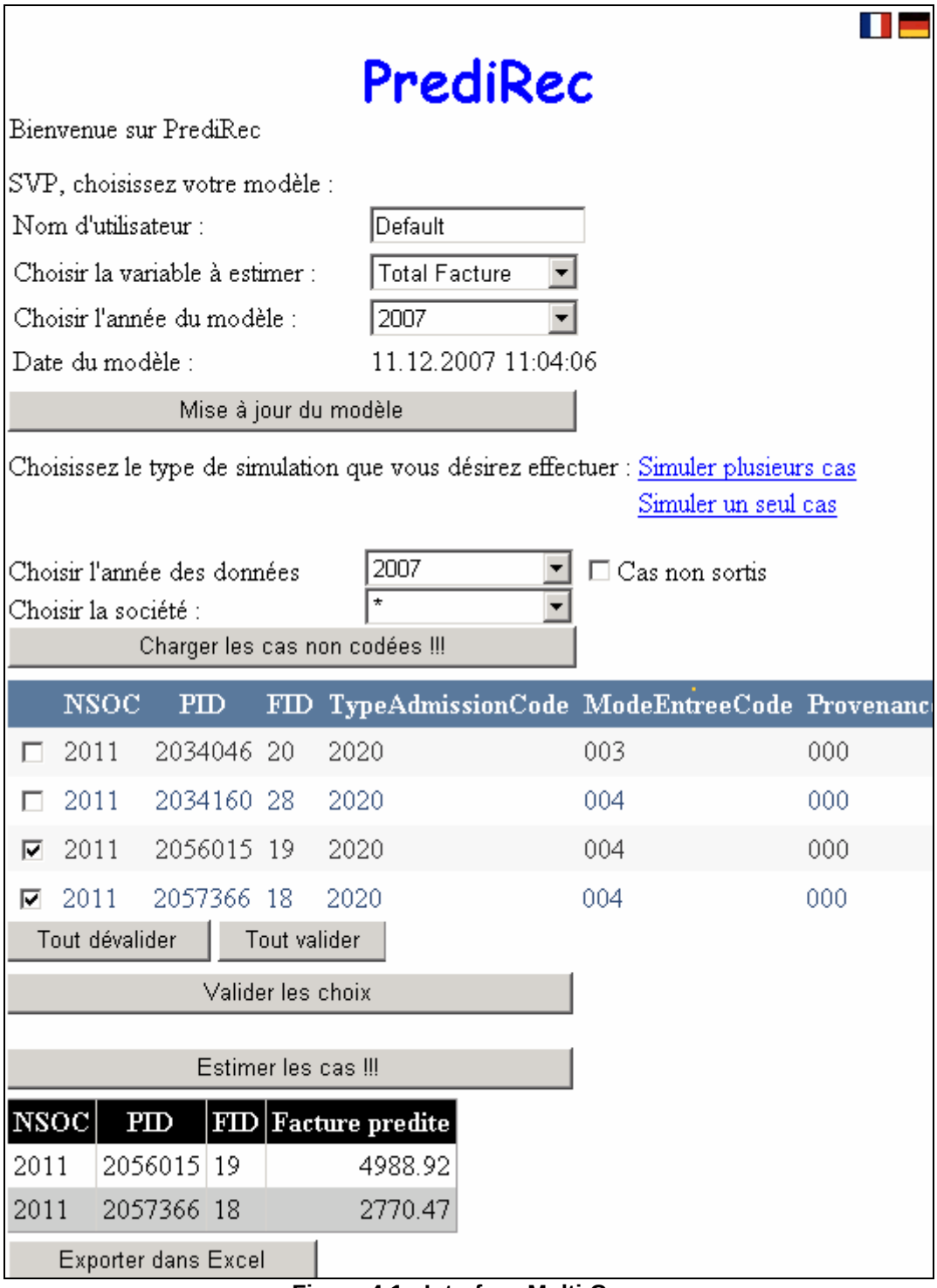


Figure 4-1 : Interface Multi-Case

4.2 Interface Single-Case

Exemple de l'interface pour la spécification d'un cas unique :

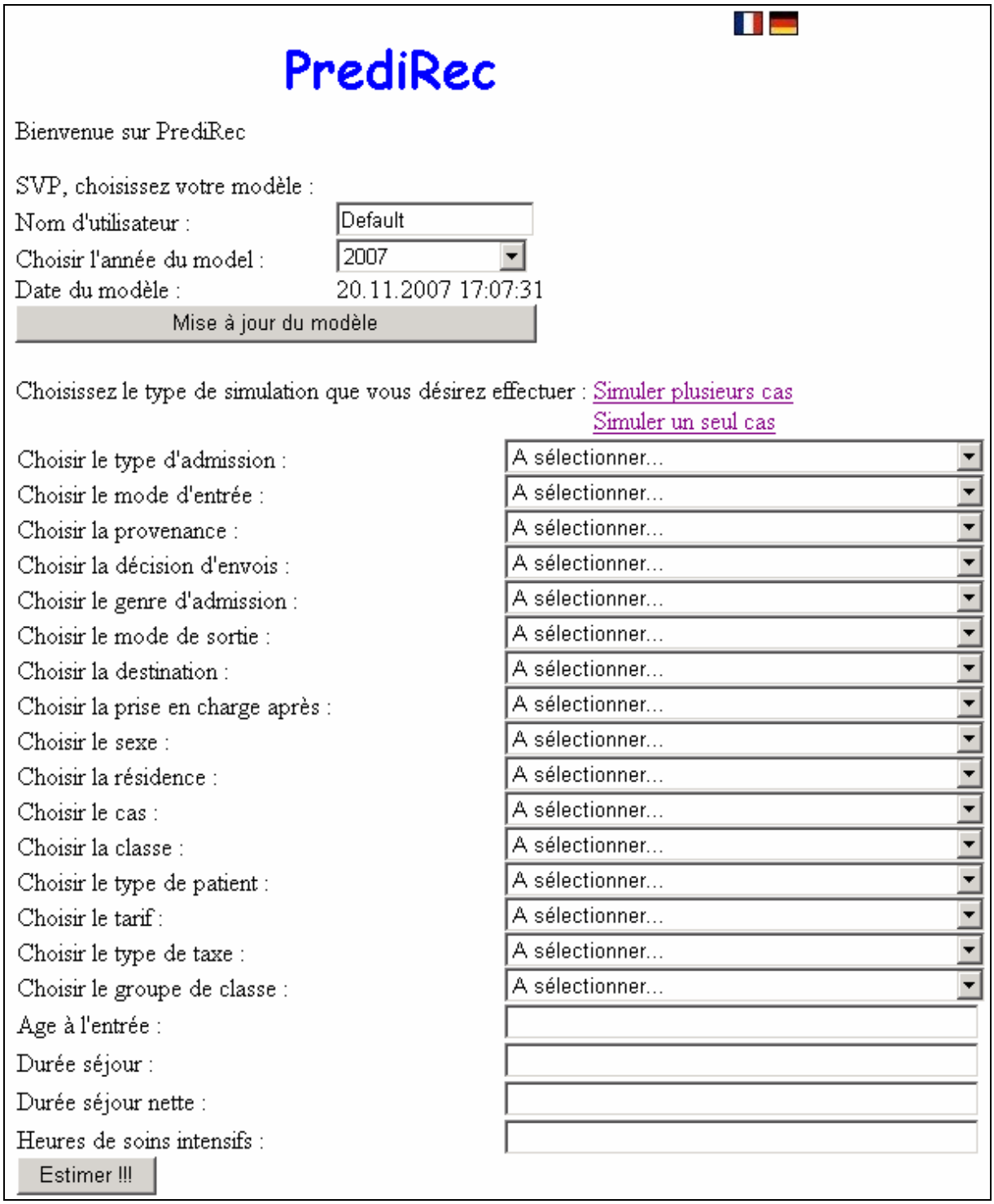


Figure 4-2 : Interface Single-Case

5 Conclusion

Le SIMAV désire mettre à disposition une solution informatique permettant à des utilisateurs précis de prédire les recettes liées à des cas non codés des hôpitaux valaisans.

Pour arriver à ce but, il a été décidé de mettre en place une solution de Data Mining : PrediRec.

L'essentiel de ce projet consiste en la recherche de modèles d'exploration de données permettant d'approximer au mieux les recettes réelles des cas non codés.
Afin de choisir un modèle d'exploration de données, un set de variables est à disposition pour les phases de test et d'apprentissage.

Un site web permet aux utilisateurs finaux de PrediRec d'utiliser les modèles d'exploration mis à leur disposition.
L'interface utilisateur de cette application dispose de fonctionnalités réduites au minimum et l'effort de développement sera consacré à la précision des résultats des estimations.

Projet SIMAV

Compte rendu d'entretien

Codificatrice

Facturation par APDRG : prédiction des recettes des cas non codés

Institut Central des Hôpitaux Valaisans
Av. Grand Champsec 86
Case postale 736
1951 Sion
Suisse

Auteur	: Giotta Mathieu	**Date de création**	: 02.10.2007
Fichier	: PrediRec - 006 EntretienUtilisateur Codificatrice	**No de version**	: V. finalisée
Etat	: Approuvé	**Dernière révision**	:
Distribution	:	**Date de distribution**	:
Publication	:		

Table des matières

1 Date de l'entretien

Mardi 2 octobre 2007

2 Personne de contact

Mme. Caroline Farmer

3 Statut de la personne de contact

Responsable des codificatrices

4 But de l'entretien

Le but de cet entretien est d'avoir connaissance de la méthodologie utilisée actuellement par les codificatrices pour coder le dossier d'un patient sorti de l'hôpital.

5 Lieu de l'entretien

L'entretien a eu lieu dans le bureau des codificatrices du site de Sion.

6 Durée de l'entretien

2 heures

7 Résultat de l'entretien

Mme. Farmer m'a expliqué la procédure ci-dessous, expliquant la méthodologie pour le codage du dossier d'un patient.

7.1 Procédure de codage

Les codificatrices reçoivent, en général, un dossier papier indiquant la sortie du patient de l'hôpital.

Dès le dossier en main, la codificatrice se connecte dans Phoenix pour contrôler que le patient est marqué comme sortis dans Phoenix.

- Si le patient n'est pas sorti, la codificatrice informe le médecin responsable du patient afin que celui-ci ferme le dossier patient électronique et le dossier est mis en attente.
- Si le dossier est fermé dans Phoenix, la codificatrice peut continuer le codage.

A la suite du contrôle Phoenix, la codificatrice commence à introduire des informations sur le dossier du patient dans Opale, plus précisément dans l'écran de saisie ctrl+F8 (Figure 7-1).

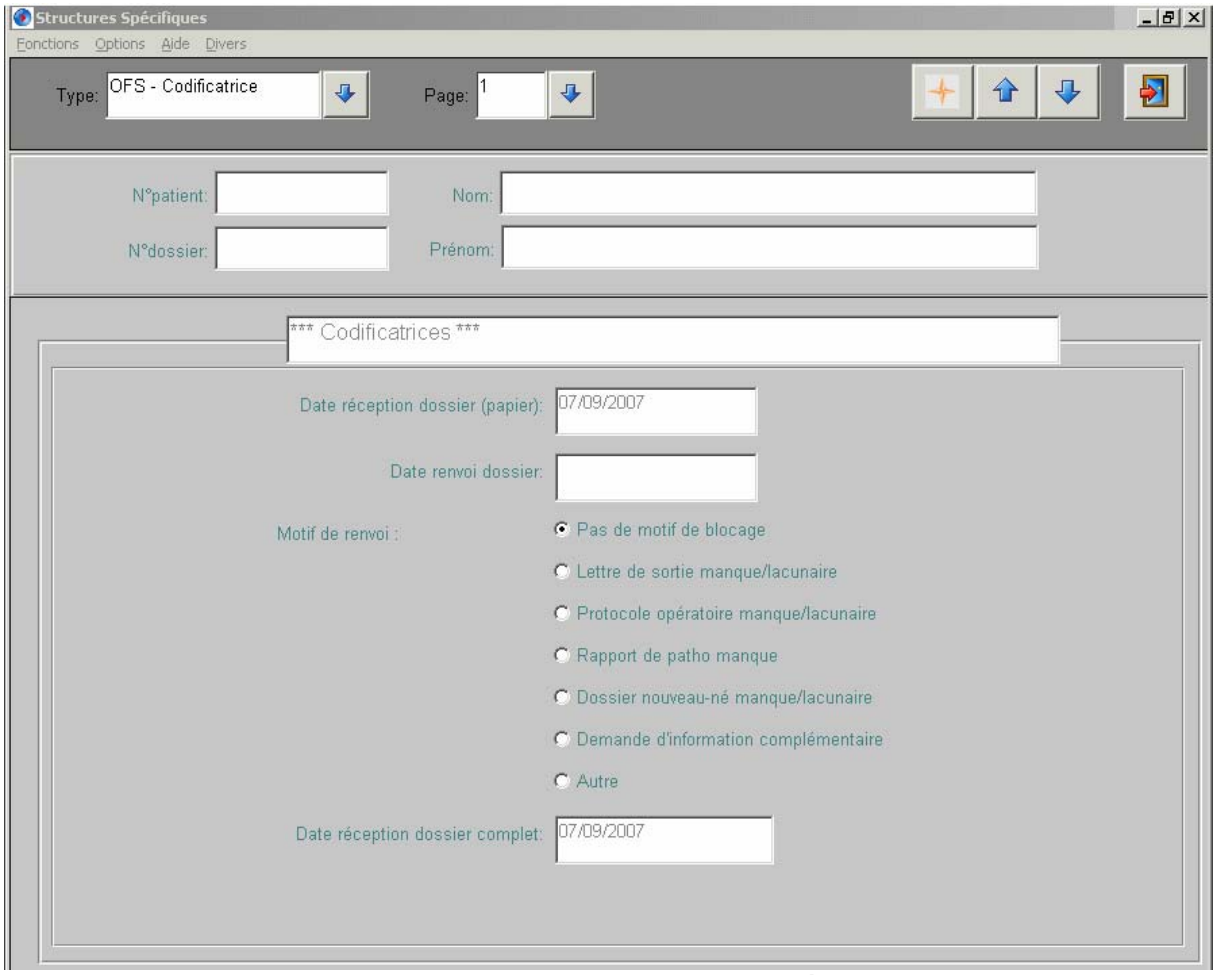


Figure 7-1 :Ecran Opale – Structures spécifiques

Dans cet écran la codificatrice contrôle que tous les éléments nécessaires au codage soient présents et, si il en manque, la codificatrice informe le médecin responsable du dossier et met le dossier en attente.

Si le dossier est complet, elle peut continuer le codage.

Une fois le dossier complet, la codificatrice doit se rendre dans l'écran Opale de saisie des données OFS (ctrl+j) ci-dessous.

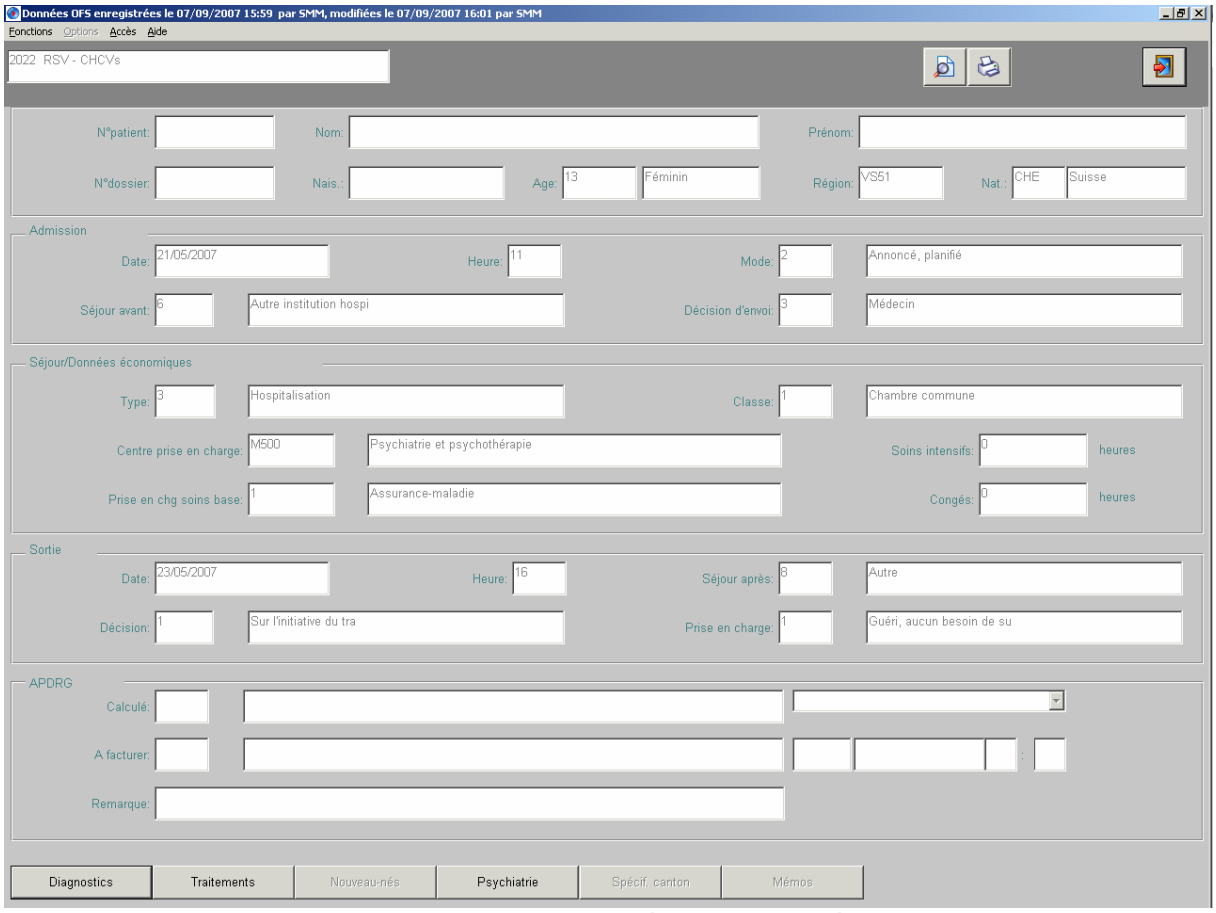


Figure 7-2 : Ecran Opale – Données OFS, données de base

Dans cet écran, la codificatrice complète certaines informations, qui peuvent manquer pour la formalisation du dossier électronique Opale en dossier OFS.

A la suite de quoi, elle cliquer sur le bouton Diagnostics, action qui fera apparaître l'écran ci-dessous

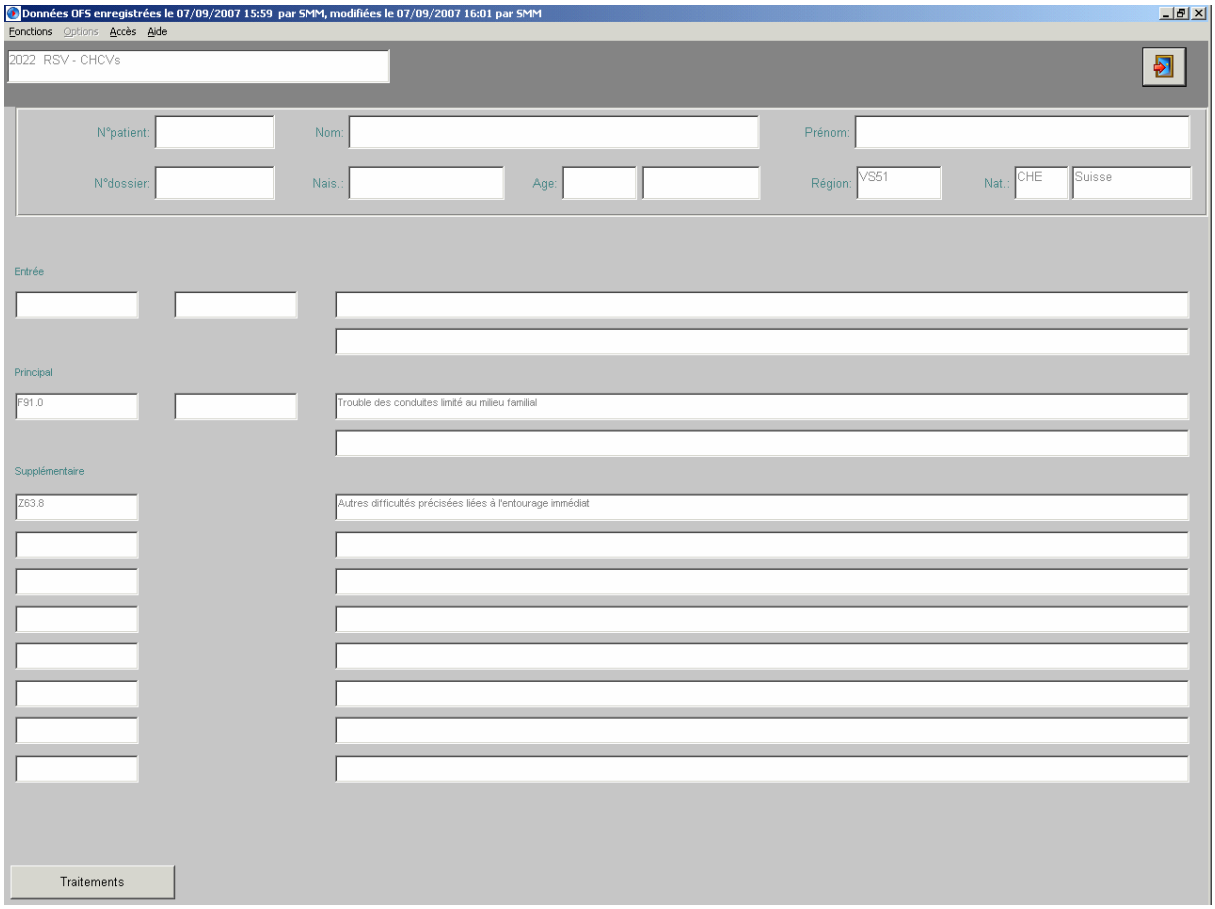


Figure 7-3 : Ecran Opale – Données OFS, codes diagnostics

Dans cet écran, la codificatrice interprètera la lettre de sortie et la traduira en différents codes de diagnostics.

Après avoir introduit les codes de diagnostics, la codificatrice doit introduire les codes de traitements, dans l'écran approprié en cliquant sur le bouton Traitements

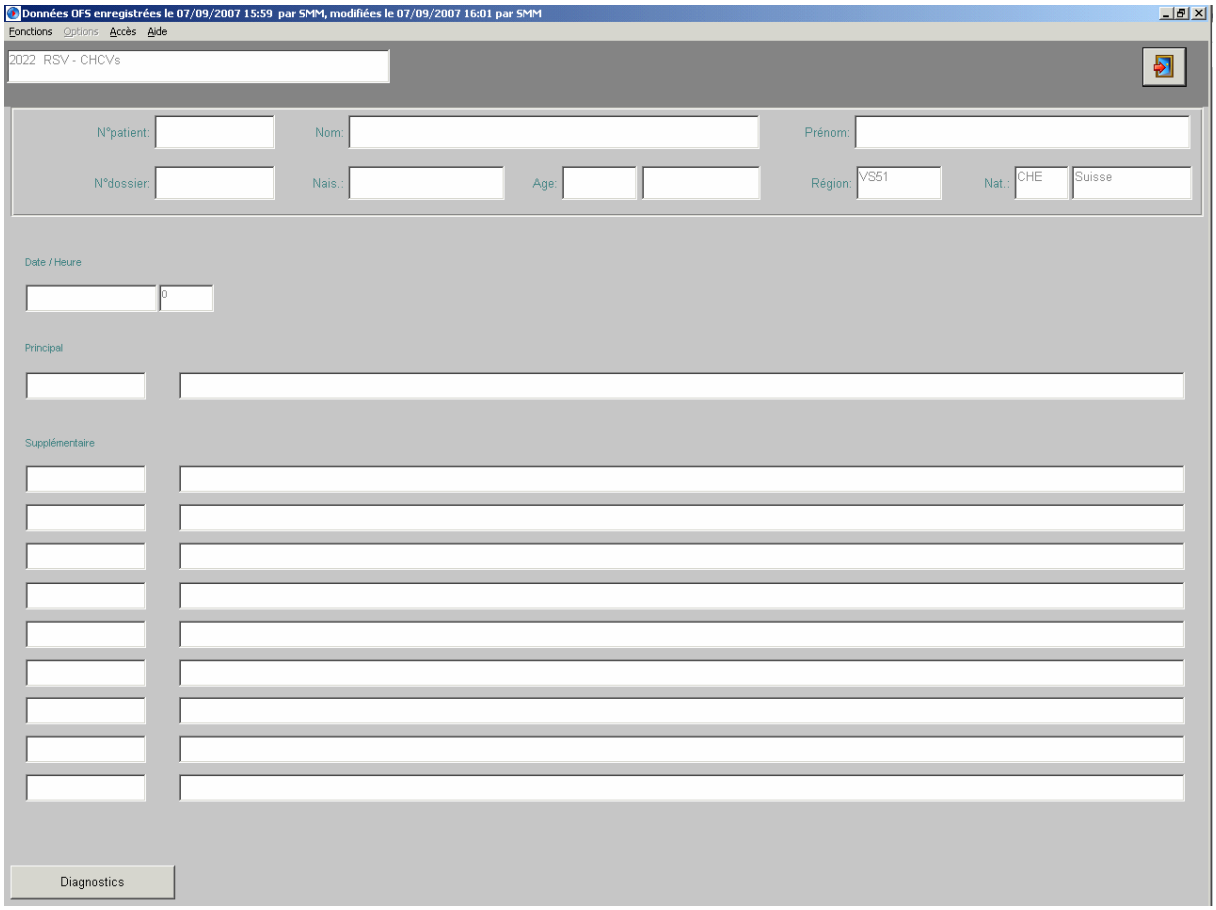


Figure 7-4 : Ecran Opale – Données OFS, codes de traitements

Une fois que les informations nécessaires sont introduites dans le système, le codage se termine simplement en cliquant sur le bouton

A ce moment là, les informations sont groupées, c'est-à-dire quelles sont reprises dans un logiciel tiers qui interprète les codes de diagnostique, les codes de traitement et les autres information de l'écran OFS pour une résulter un code APDRG ainsi qu'un CW, qui détermine le montant de la facture finale ➔ entretien avec les facturistes.

Projet SIMAV

Compte rendu d'entretien

M. Thomas Werlen

Facturation par APDRG : prédiction des recettes des cas non codés

Institut Central des Hôpitaux Valaisans
Av. Grand Champsec 86
Case postale 736
1951 Sion
Suisse

Auteur	: Mathieu Giotta	**Date de création**	: 10.10.2007
Fichier	: PrediRec - 007 EntretienUtilisateur Werlen	**No de version**	: V. finalisée
Etat	: Approuvé	**Dernière révision**	:
Distribution	:	**Date de distribution**	:
Publication	:		

Table des matières

1 Date de l'entretien

Mercredi 10 octobre 2007

2 Personne de contact

M. Thomas Werlen

3 Statut de la personne de contact

Responsable administratif, finances et qualité du SZO

4 But de l'entretien

Le but de cet entretien est de prendre connaissance de la méthode utilisée jusqu'à aujourd'hui dans les différents hôpitaux pour évaluer les cas non codés.

5 Lieu de l'entretien

L'entretien a eu lieu au SZO, site de Brig.

6 Durée de l'entretien

2 heures

7 Résultat de l'entretien

M. Werlen m'a donc expliqué la méthode qu'il utilise actuellement pour provisionner les recettes des cas non codés.
Je décris cette procédure ci-dessous.

7.1 Procédure de prévision des recettes

En début de chaque année, ainsi qu'en cours d'année lors des bouclements trimestriels, M. Werlen exécute une liste paramétrée dans le système administratif Opale qui permet d'extraire dans Excel tous les cas, pour une période donnée, avec leur APDRG.

Dans ce classeur Excel, les cas non codés possèdent un code APDRG de 0, code qui n'existe pas dans la table de référence des APDRGs.

Depuis ce fichier, M. Werlen extrait les cas où l'APDRG égal 0.

Ensuite, toujours dans le fichier Excel, il calcule le CW moyen pour chaque type de cas (un type de cas peut être soit de la chirurgie, de l'orthopédie, de la médecine, etc. Il existe actuellement plus d'une septantaine de type de cas).

Par après, il attribue à chacun des cas non codés le CW moyen correspondant à son type de cas et multiplie le CW moyen par la valeur du point APDRG de l'exercice correspondant (chaque année, la valeur du point est réévaluée.).

Une fois que ces opérations sont faites, M. Werlen saisit les provisions dans OPALE, dans des différents comptes transitoires, selon le type de cas du cas non codé.

Projet SIMAV

Compte rendu d'entretien

M. Bellani et M. Levrand

Facturation par APDRG : prédiction des recettes des cas non codés

Institut Central des Hôpitaux Valaisans
Av. Grand Champsec 86
Case postale 736
1951 Sion
Suisse

Auteur	: Mathieu Giotta	**Date de création**	: 07.11.2007
Fichier	: PrediRec - 008 EntretienUtilisateur Facturistes	**No de version**	: V. finalisée
Etat	: Approuvé	**Dernière révision**	:
Distribution	:	**Date de distribution**	:
Publication	:		

Table des matières

1 Date de l'entretien

Mercredi 07 novembre 2007

2 Personnes de contact

M. Marius Levrand
M. Gilbert Bellani

3 Statut des personnes de contact

Responsable de la codification ainsi que de la facturation des sites du CHCVs Sion et Martigny.

4 But de l'entretien

Le but de cet entretien est de comprendre comment sont facturés les cas codés des hôpitaux valaisans et aussi d'essayer de définir quelles sont les variables potentiellement utilisables des systèmes sources (Opale et Phoenix).

5 Lieu de l'entretien

L'entretien a eu lieu au CHCVs, site de Sion.

6 Durée de l'entretien

3 heures

7 Résultats de l'entretien

7.1 Facturation

Les facturistes reprennent le Cost-Weight pondéré calculé par les codificatrices et lui attribue une valeur du point qui dépend :

- de l'année de sortie du patient de l'hôpital ;
- du type de tarif est attribué au patient (Hospitalisation commune, hospitalisation privée, Hospitalisation commune hors convention, etc.).

Une fois la valeur du point sélectionnée, ils peuvent effectuer la facture et l'envoyer.

7.2 Variables potentiellement utilisables

Les facturistes suggèrent d'utiliser les variables suivantes pour la mise en place du modèle d'analyse :

- la résidence pour convention ;
- l'âge à l'entrée ;
- le type de cas ;
- la classe d'hospitalisation ;
- le type de patient ;
- le tarif ;
- le service ;
- l'unité ;
- le médecin traitant ;
- la durée de séjour ;

- la durée de séjour nette ;
- les heures de soins intensifs ;
- la liste des prestations Tarmed, leurs points et leurs valeurs.

Ils savent que certaines de ces variables influencent le calcul du Cost-Weight pondéré et les autres influencent la valeur du point à attribuer. Certaine d'entre elles permettent simplement d'évaluer la lourdeur des cas.

Projet SIMAV

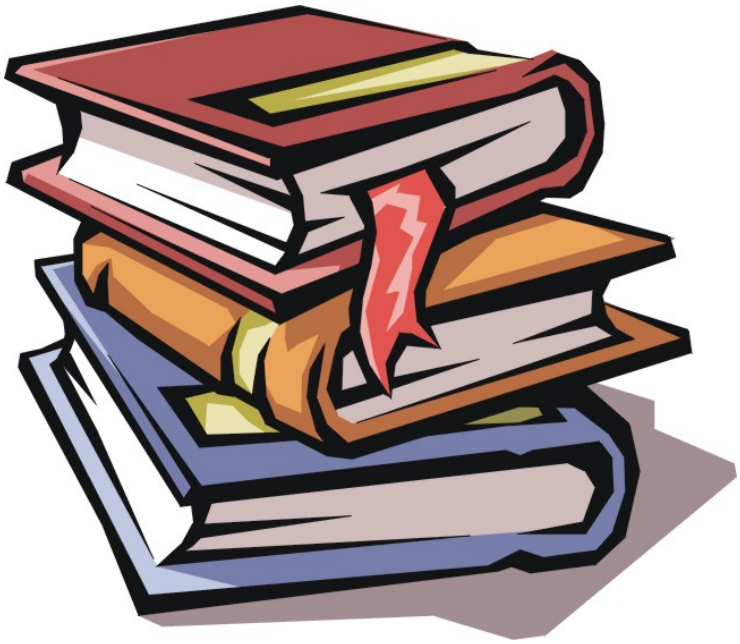
Mode d'emploi

Facturation par APDRG : prédiction des recettes des cas non codés

Institut Central des Hôpitaux Valaisans
Av. Grand Champsec 86
Case postale 736
1951 Sion
Suisse

Auteur	: Mathieu Giotta	**Date de création**	: 15.12.2007
Fichier	: PrediRec - 009 Mode d'emploi	**No de version**	: V finale
Etat	:	**Dernière révision**	:
Distribution	:	**Date de distribution**	:
Publication	:		

Table des matières



1 Introduction

A la fin de chaque année ainsi que lors des bouclements comptables trimestriels, il est nécessaire de provisionner les cas sortis des hôpitaux et qui ne sont pas encore facturés.

Actuellement, les hôpitaux calculent leurs provisions pour le secteur somatique aigu de manière empirique sur la base d'un montant moyen forfaitaire.
Cette solution ne donne pas entière satisfaction, car elle ne tient pas compte de la lourdeur des cas et, de plus, cette méthode est assez fastidieuse.

Il a donc été décidé de mettre en place une solution qui permet d'estimer au plus juste ces cas.

Cette solution se nomme « PrediRec » et est la concaténation des termes « Prédiction » et « Recette ».

1.1 Rappel de la méthode RSV

En début de chaque année, ainsi qu'en cours d'année lors des bouclements trimestriels, les comptables exécutent une liste paramétrée dans le système administratif Opale qui permet d'extraire dans Excel tous les cas, pour une période donnée, avec leur APDRG.

Dans ce classeur Excel, les cas non codés possèdent un code APDRG de 0, code qui n'existe pas dans la table de référence des APDRGs.

Depuis ce fichier, les comptables extraient les cas où l'APDRG égal 0.

Ensuite, dans le fichier Excel, ils calculent le CW moyen pour chaque type de cas (un type de cas peut être soit de la chirurgie, de l'orthopédie, de la médecine, etc. Il existe actuellement plus d'une septantaine de type de cas).

Par après, ils attribuent à chacun des cas non codés le CW moyen correspondant à son type de cas et multiplie le CW moyen par la valeur du point APDRG de l'exercice correspondant (chaque année, la valeur du point est réévaluée.).

Une fois que ces opérations sont effectuées, les comptables introduisent ces estimations en tant que provisions dans OPALE, à travers différents comptes transitoires, selon le type de cas du cas non codé.

2 Description générale de PrediRec

PrediRec est une solution de Data Mining permettant de simuler les recettes liées aux cas non codés des hôpitaux valaisans et pour lesquels les comptables doivent effectuer des prévisions.

2.1 Vue d'ensemble des fonctionnalités

L'application Web PrediRec et en mesure de proposer les fonctionnalités suivantes :

Choix de la variable à estimer
L'utilisateur doit sélectionner quelle est la variable qu'il désire estimer :

- soit le « CW pondéré » ;
- soit le « Total Facture ».

Mise à jour des modèles d'exploration de données
Par cette opération, l'utilisateur effectue la phase d'apprentissage nécessaire à l'application de Data Mining.
Chaque utilisateur crée son propre modèle, c'est-à-dire qu'un utilisateur « x » peut mettre à jour son modèle de données pendant que l'utilisateur « y » travaille sur le sien.
De cette manière, un utilisateur peut évaluer ses cas non codés sans risquer d'être déstabilisé par une mise à jour impromptue du modèle.

Choix des cas non codés
L'utilisateur final peut sélectionner les cas non codés qu'il désire estimer. PrediRec se connecte à la base de données du Data Warehouse pour en extraire les informations relatives aux cas choisis.

Estimation des cas choisis
La principale fonctionnalité de PrediRec repose sur l'estimation des cas non codés choisis. Par cette action, l'utilisateur soumet ces cas au moteur de Data Mining qui les lui retourne, complétés par son évaluation.

Exportation des cas simulés dans MS Excel
A la suite d'une simulation de cas non codés, l'utilisateur peut exporter les résultats dans MS Excel.

Simulation d'un cas fictif
L'utilisateur de PrediRec a la possibilité de créer un cas fictif dans le but, par exemple, d'effectuer un devis avant une hospitalisation.

Choix de la langue de l'application Web
Etant donné cette application peut être utilisée par des germanophones, l'utilisateur est libre de choisir la langue des textes affichés.

3 Mode d'emploi

3.1 Connexion au système

L'application PrediRec est intégrée au portail du Data Warehouse, c'est-à-dire « InfoView » (Figure 3-1). La connexion au sein du système ce fait donc via votre login qui vous a été communiqué par l'équipe du Data Warehouse.

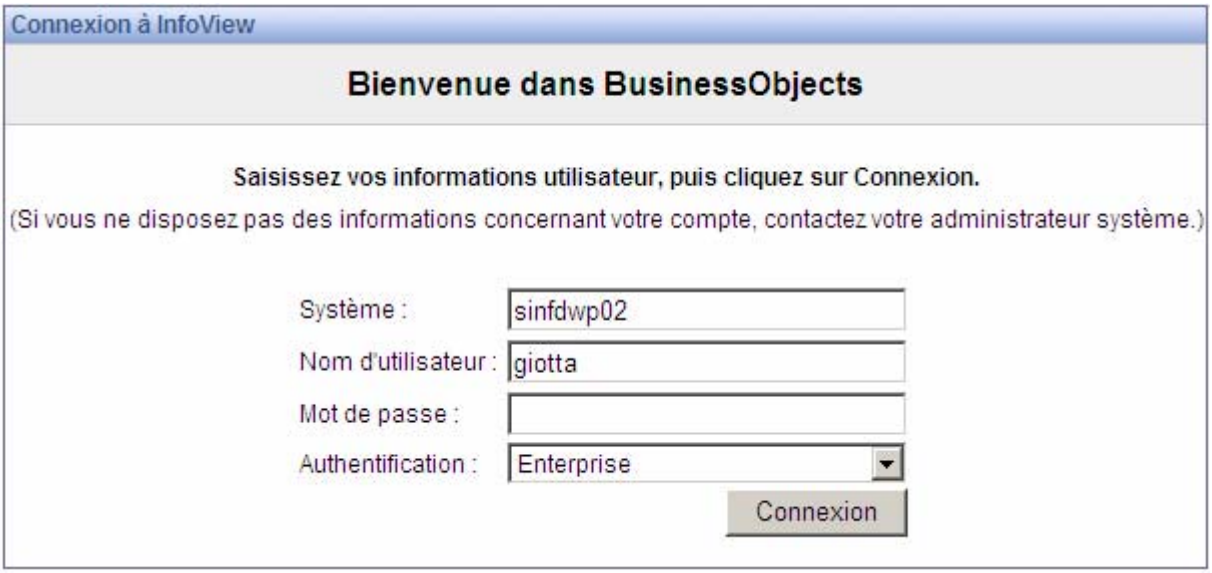


Figure 3-1 : Ecran de connexion à Infoview

Une fois connecté au système, vous avez la possibilité de choisir :

- de simuler plusieurs cas ;
- de simuler un cas fictif.

à l'aide de lien hypertexte(Figure 3-2).

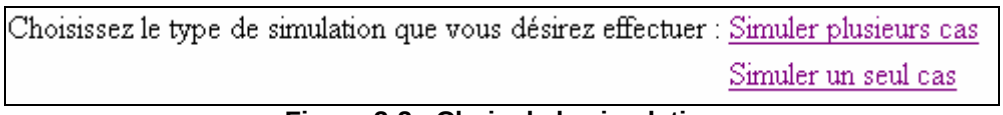


Figure 3-2 : Choix de la simulation

Il est aussi possible de changer la langue des textes affichés de PrediRec en la choisissant selon le drapeau dans le coin en haut à droite (Figure 3-3).

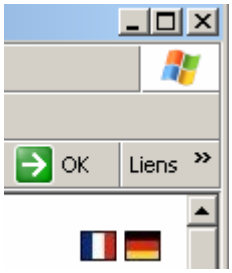


Figure 3-3 : Choix de la langue

3.2 Simulation de plusieurs cas

PrediRec permet l'extraction depuis le Data Warehouse de tous les cas non codés des hôpitaux valaisans.

Nota-Bene
Etant donné que le chargement du Data Warehouse se fait la nuit, lorsque vous effectuez des prédictions, les cas codés durant la journée ne le sont pas dans le DW (un jour de retard).

3.2.1 Choix du modèle d'analyse

Une fois l'application PrediRec à l'écran, la première étape consiste à changer le nom d'utilisateur (Figure 3-4). Si cette étape est omise et que le nom d'utilisateur reste « Default », l'application de ne fonctionnera pas.

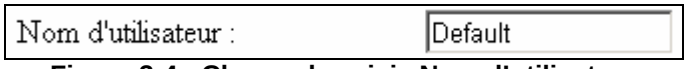


Figure 3-4 : Champ de saisie Nom d'utilisateur

Une fois le nom saisi, il est nécessaire de choisir quelle est la variable que vous désirez prédire depuis la liste déroulante « Choisir la variable à estimer » (Figure 3-5) ainsi que l'année à laquelle correspond le modèle d'analyse que vous souhaitez utiliser (Figure 3-6). Au moment où la variable à prédire et l'année du modèle sont sélectionnées, PrediRec affiche la date de la dernière mise à jour du modèle (Figure 3-7), et si dans l'éventualité où vous n'avez jamais utilisé de modèle d'analyse pour cette année et cette variable choisies, PrediRec effectue la mise à jour du modèle automatiquement.

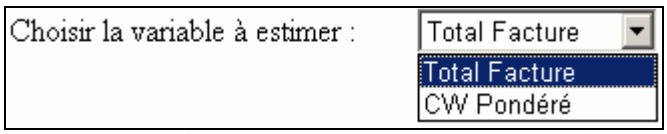


Figure 3-5 : Liste déroulante « Choix de la variable à estimer »

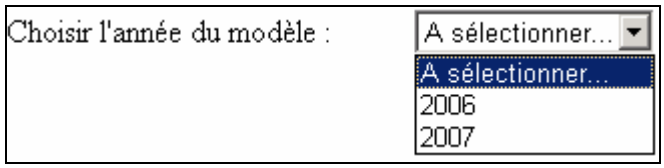


Figure 3-6 : Liste déroulante « Choix de l'année du modèle »

Figure 3-7 : Date du dernier rafraîchissement du modèle

Nous conseillons de procéder à la mise à jour du modèle d'analyse à chaque utilisation de PrediRec. Nous n'avons pas activé cette fonctionnalité car, vous pouvez devoir effectuer une simulation durant plusieurs jours et il est nécessaire qu'à chacune des simulations vous retrouviez les mêmes résultats.

La mise à jour du méthode se fait simplement en cliquant sur le bouton « Mise à jour du modèle » (Figure 3-8).
Attention : lors du clic sur ce bouton, vous mettez à jour uniquement le modèle de la variable sélectionnée. Si vous désirez mettre à jour les deux modèles, il est nécessaire de changer la variable et de recliquer sur le bouton.
Une mise jour d'un modèle dure en général moins d'une minute.

Figure 3-8 : Bouton « Mise à jour du modèle »

Durant la mise à jour du modèle d'analyse, PrediRec ne donne aucun signe d'activité, mais Internet Explorer 6 affiche une barre de progression dans sa barre d'état (Figure 3-9).
A la fin de la mise à jour du modèle, la Figure 3-7 se rafraîchit et affiche la date et l'heure actuel du système.

Figure 3-9 : Barre de progression d'Internet Explorer 6

3.2.2 Choix des cas non codés

Le choix des cas non codés se fait simplement à l'aide de deux listes déroulantes et d'une case à cocher (Figure 3-10).

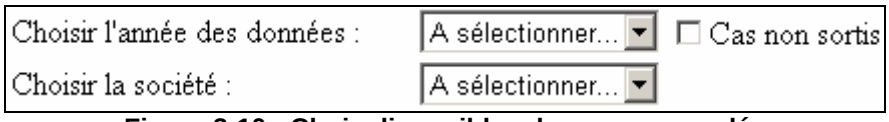


Figure 3-10 : Choix disponibles des cas non codés

A partir de la première liste déroulante(« Choisir l'année de données » (Figure 3-11), il est nécessaire de choisir l'année de sortie des cas non codés.

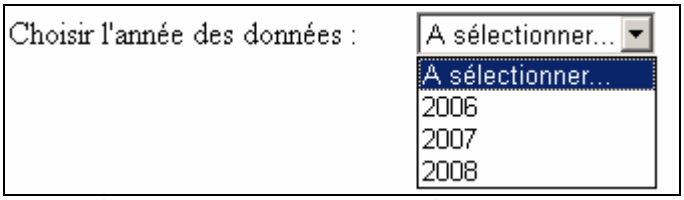


Figure 3-11 : Liste déroulante « Choix de l'année de sortie des cas non codés »

Etant donné qu'il peut s'avérer parfois nécessaire d'estimer les cas non sortis des hôpitaux, vous pouvez les sélectionner en activant la case à cocher « Cas non sortis » (Figure 3-12).

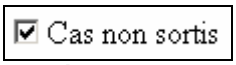


Figure 3-12 : Case à cocher « Cas non sortis »

Depuis la seconde liste déroulante (Figure 3-13), vous devez choisir quelle est la société à partir de laquelle PrediRec doit simuler les cas non codés. Si vous désirez effectuer une simulation des cas non codés des trois centres hospitaliers du RSV, vous pouvez y sélectionner l'astérisque (*) et PrediRec affiche tous les cas non codés du canton.

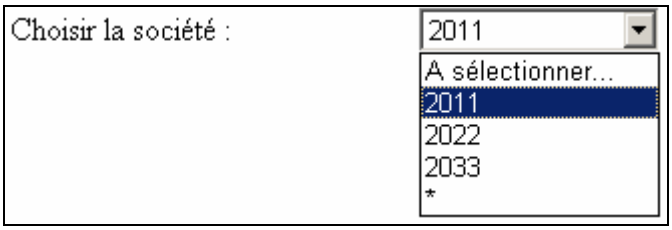


Figure 3-13 : Liste déroulante « Choix de la société »

Nota-Bene : Il possible d'activer uniquement la case à cocher « Cas non sortis » pour effectuer une prévision seulement des cas non sortis des hôpitaux valaisans.

Une fois les choix effectués, il faut cliquer sur le bouton « Charger les cas non codés » (Figure 3-14) afin que PrediRec affiche les cas non codés correspondant à votre sélection.

Figure 3-14 : Bouton « Charger les cas non codés !!! »

Si PrediRec trouve des cas non codés respectant vos choix, il les affiche dans un tableau (Figure 3-15) avec quelques informations complémentaires. Vous avez la possibilité de choisir parmi tous les cas non codés affichés, quels sont ceux que vous désirez estimer en activant ou en désactivant la case à cocher se trouvant en regard de chaque cas non codé. Une case activée signifie que le cas doit être estimé.

Par défaut, tous les cas non codés sont sélectionnés, mais il existe deux boutons qui permettent d'activer (Figure 3-16) ou de désactiver (Figure 3-17) tous les cas non codés.

	NSOC	PID	FID	Age à l'entrée	Cas	Classe
☐	2011	2034046	20	69	1610	01
☐	2011	2034160	28	54	1610	01
☑	2011	2056015	19	41	1610	01
☑	2011	2057366	18	51	1610	01

Figure 3-15 : Tableau des cas non codés

Figure 3-16 : Bouton « Tout dévalider »

Figure 3-17 : Bouton « Tout valider »

Une fois le choix des cas effectué, il ne vous reste plus qu'a cliquer sur le bouton « Valider les choix.. » (Figure 3-18).

Figure 3-18 : Bouton « Valider les choix… »

3.2.3 Estimation des cas non codés

La dernière étape consiste à exécuter l'estimation des cas non codés. Pour le faire, il suffit de cliquer sur le bouton « Estimer les cas !!! » (Figure 3-19).

Figure 3-19 : Bouton « Estimer les cas !!! »

A la suite de cette action, PrediRec affiche dans un nouveau tableau (Figure 3-20) l'estimation des cas non codés qui lui ont été indiqués à l'étape précédentes « 3.2.2 - Choix des cas non codés ».

NSOC	PID	FID	Prédiction
2011	2056015	19	5002.67
2011	2057366	18	2654.06

Figure 3-20 : Tableau des résultats

Comme vous pouvez l'observer, à la Figure 3-15 nous avions désélectionner deux cas et donc ceux-ci ne sont pas affichés dans la Figure 3-20.

3.2.4 Exportation des résultats dans MS Excel

Une fois l'estimation exécutée, PrediRec permet d'exporter ses résultats dans MS Excel si d'autres manipulations de ces résultats son nécessaire.

Pour exporter les prédictions, il suffit de cliquer sur le bouton « Exporter dans Excel » (Figure 3-21), de cliquer dans la boite de dialogue affichée à l'écran (Figure 3-22) sur le bouton « Enregistrer » et, dans la fenêtre suivante (Figure 3-23), indiquer quel est le nom que l'on souhaite donner au fichier et indiquer où est-ce que le fichier doit être enregistré et valider le tout en cliquant à nouveau sur « Enregistrer ».

Figure 3-21 : Bouton « Exporter dans Excel »

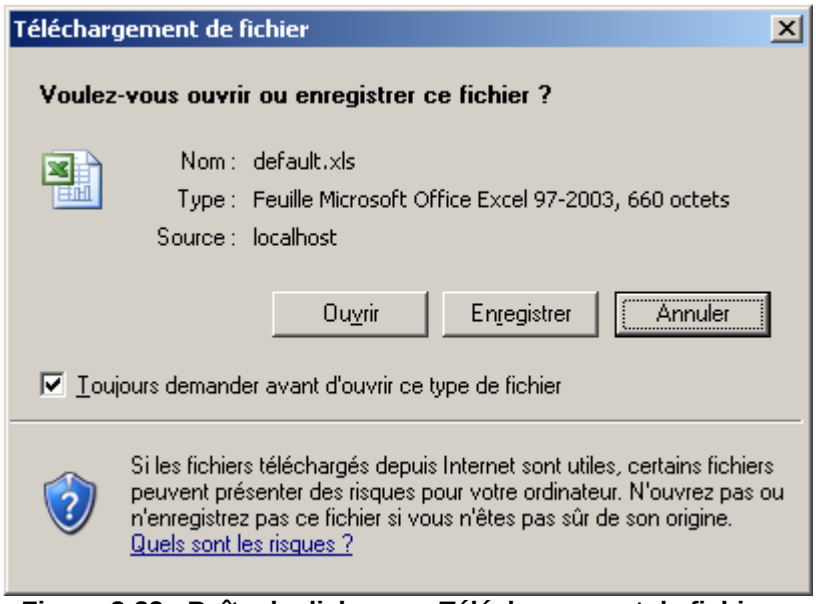


Figure 3-22 : Boîte de dialogue « Téléchargement de fichier »

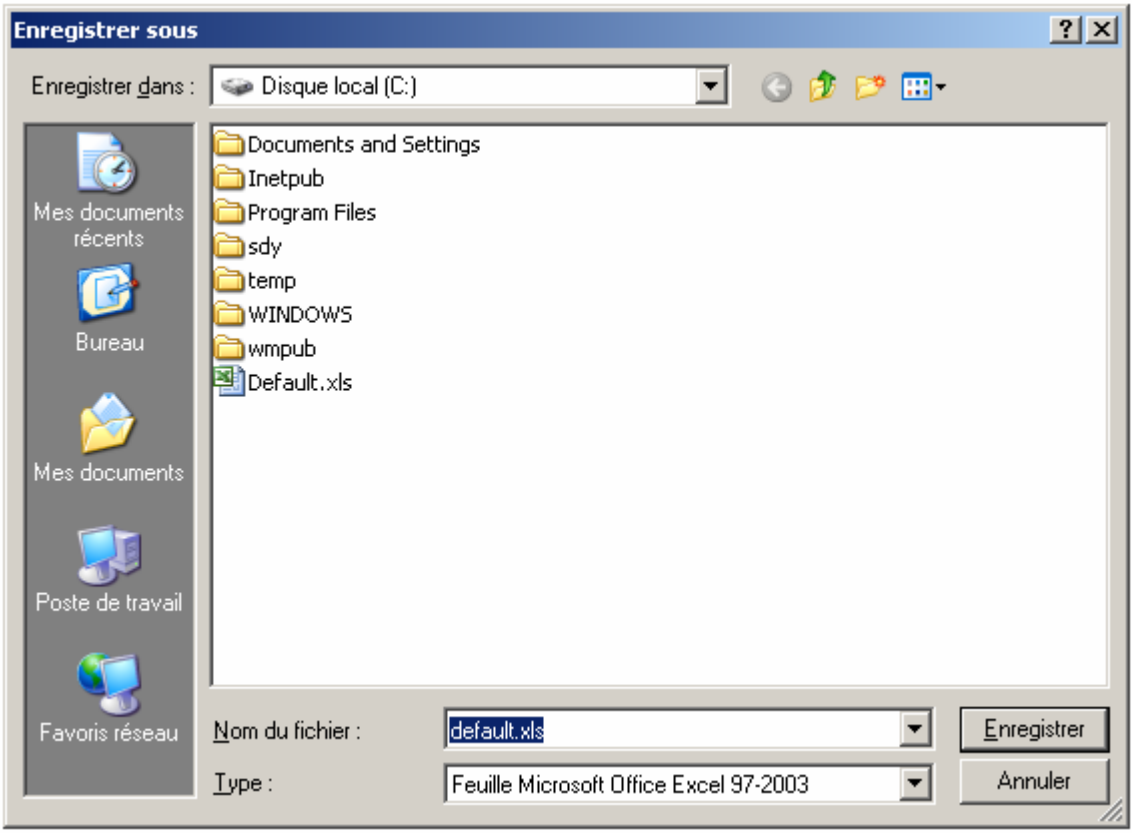


Figure 3-23 : Boîte de dialogue « Enregistrer sous »

3.3 Simulation de cas fictif

3.3.1 Choix du modèle d'analyse

Comme pour la simulation de plusieurs cas, il est nécessaire de choisir le modèle d'analyse.

Etant donné que cette procédure vous a déjà été expliquée, nous ne la réexpliquons pas et nous vous laissons consulter le chapitre « 3.2.1 - Choix du modèle d'analyse ».

3.3.2 Saisie d'un cas fictif

La première étape consiste à la saisie de l'âge du patient (Figure 3-24).

Figure 3-24 : Champ de saisie « Age à l'entrée »

Ensuite, il faut sélectionner dans la liste déroulante « Choisir le cas » (Figure 3-25) quel est le type de cas du cas fictif. Cette liste comporte les différents types de cas somatiques aigus disponible dans Opale :

- 0299 - Cardiologie n/cantonalisée ;
- 0300 - Cardiologie cantonalisée ;
- 0301 - Chirurgie cardiaque cantonalisée ;
- 0302 - Chirurgie du dos (Neurochirurge) ;
- 0303 - Chirurgie Générale ;
- 0304 - Chirurgie maxillo-faciale ;
- 0306 - Chirurgie pédiatrique ;
- 0307 - Chirurgie Plastique/Reconstructive ;
- 0308 - Chirurgie thoracique ;
- 0309 - Chirurgie vasculaire ;
- 0310 - Chirurgie Esthétique ;
- 0311 - Chirurgie cardiaque n/cantonalisée ;
- 0402 - Dentaire ;
- 0700 - Gastro-entérologie ;
- 0710 - Gynécologie ;
- 1201 - Lithotripsie ;
- 1300 - Médecine ;
- 1302 - Médecine / Oncologie ;
- 1400 - Néonatologie ;
- 1401 - Neurochg. Spécialisée ;
- 1402 - Nurserie ;
- 1405 - Neurochg. Générale ;
- 1406 - Neurologie ;
- 1500 - Maternité-Obst. ;
- 1501 - Oncologie Lourde ;
- 1502 - Ophtalmologie ;
- 1503 - ORL ;
- 1504 - Orthopédie ;
- 1600 - Pédiatrie ;
- 1603 - Pneumologie ;
- 1805 - Radio-oncologie ;
- 2000 - Traumatologie orthopédique ;
- 2100 - Urologie.

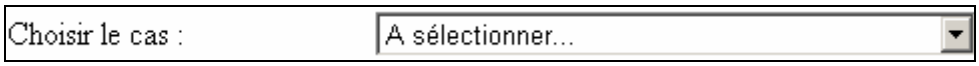


Figure 3-25 : Liste déroulante « Choisir le cas »

Après la sélection du type de cas, il est nécessaire de sélectionner quelle est la classe du séjour du cas fictif à l'aide de la liste déroulante « Choisir la classe » (Figure 3-26) :

- 01 - C - Classe Commune ;
- 02 - P - Classe Privée ;
- 10 - I - Chambre Individuelle.

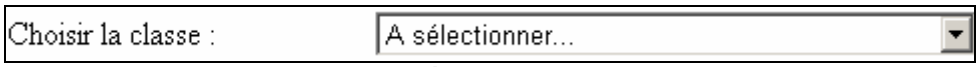


Figure 3-26 : Liste déroulante « Choisir la classe »

A la suite de la sélection de la classe du séjour, il faut spécifier quel est le groupe de classe du séjour, via la liste déroulante « Choisir le groupe de classe » (Figure 3-27):

- 01 - Classe commune ;
- 02 - Classe privée.

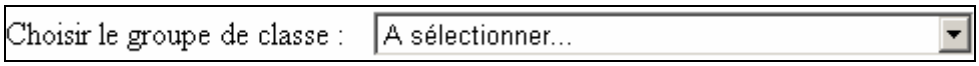


Figure 3-27 : Liste déroulante « Choisir le groupe de classe »

Ensuite, il faut spécifier quel est le sexe du cas fictif depuis la liste déroulante « Choisir le sexe » (Figure 3-28) :

- 01 - Masculin ;
- 02 - Féminin.

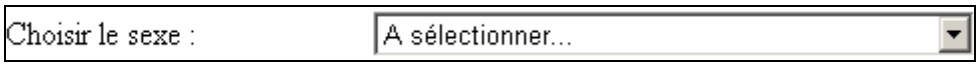


Figure 3-28 : Liste déroulante « Choisir le sexe »

Après le choix du sexe, il faut sélectionner quel est le tarif lui correspondant, via la liste déroulante « Choisir le tarif » (Figure 3-29) :

- A1 - Hosp. C - Ass. Soc. Fédérales ;
- A2 - Hosp. P - Ass. Soc. Fédérales ;
- B1 - Hosp. C - Accords Bilatéraux ;
- B2 - Hosp. P - Accords Bilatéraux ;
- C1 - Hosp. C - H.C. Commodité;
- C2 - Hosp. P - H.C. Commodité;
- D1 - Hosp. C - JU/FR CVP ;
- E1 - Hosp. C - Etranger ;
- I1 - Hosp. C - Intercantonale Urgence ;
- I2 - Hosp. P - Intercantonale Urgence ;
- M1 - Hosp. C - VS ;
- M2 - Hosp. P - VS ;
- U1 - Hosp. C - H.C. Urgence ;
- U2 - Hosp. P - H.C. Urgence.

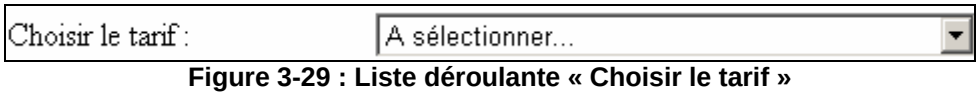


Figure 3-29 : Liste déroulante « Choisir le tarif »

Après le choix du tarif, il faut sélectionner quel est son type de patient, via la liste déroulante « Choisir le type de patient » (Figure 3-30) :

- 10 - Caisse-maladie VS ;
- 11 - Caisse-maladie Hors-canton (Urgences) ;
- 12 - Caisse-maladie Hors-canton (Commodité) ;
- 20 - Aide Sociale ;

- 28 - Caisse-maladie Jura ;
- 30 - SUVA / LAA ;
- 31 - A.I. ;
- 32 - A.M.F. ;
- 41 - Accords bilatéraux (UE).

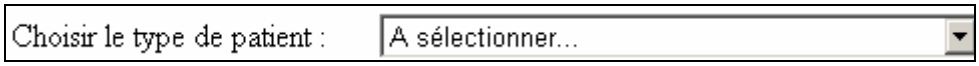


Figure 3-30 : Liste déroulante « Choisir le type de patient »

Pour terminer, il nous reste encore à saisir certaines informations :

- la durée de séjour (Figure 3-31) ;
- la durée de séjour nette (Figure 3-32) ;
- les heures de soins intensifs(Figure 3-33).

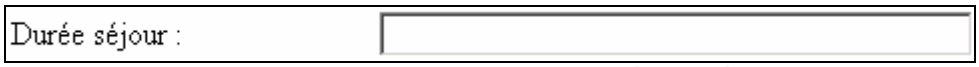


Figure 3-31 : Champ de saisie « Durée de séjour »

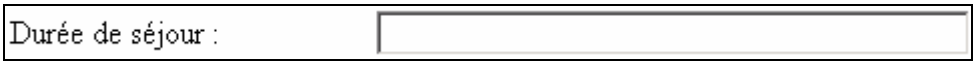


Figure 3-32 : Champ de saisie « Durée de séjour nette »

Figure 3-33 : Champs de saisie « Heures de soins intensifs »

Une fois toutes ces informations saisies ou sélectionnées, il suffit de cliquer sur le bouton « Estimer !!! » (Figure 3-34) afin que PrediRec simule le cas fictif.

Figure 3-34 : Bouton « Estimer !!! »

Le résultat de l'estimation est affiché en dessous du bouton « Estimer !!! »(Figure 3-35).

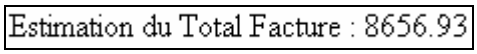


Figure 3-35 : Champ du résultat

4 Divers

Avant d'effectuer une simulation, PrediRec fait un contrôle de vos saisies ou de vos sélections et vous informe si vous avez omis d'indiquer une valeur.

Par exemple, si le nom d'utilisateur n'est pas modifié, PrediRec ne s'exécutera pas et vous informera du problème.

Dans la page « Simuler un cas fictif », si une des listes de variables n'est pas changée par rapport à sa valeur pas défaut, PrediRec vous le fait remarquer et n'effectue pas la simulation.

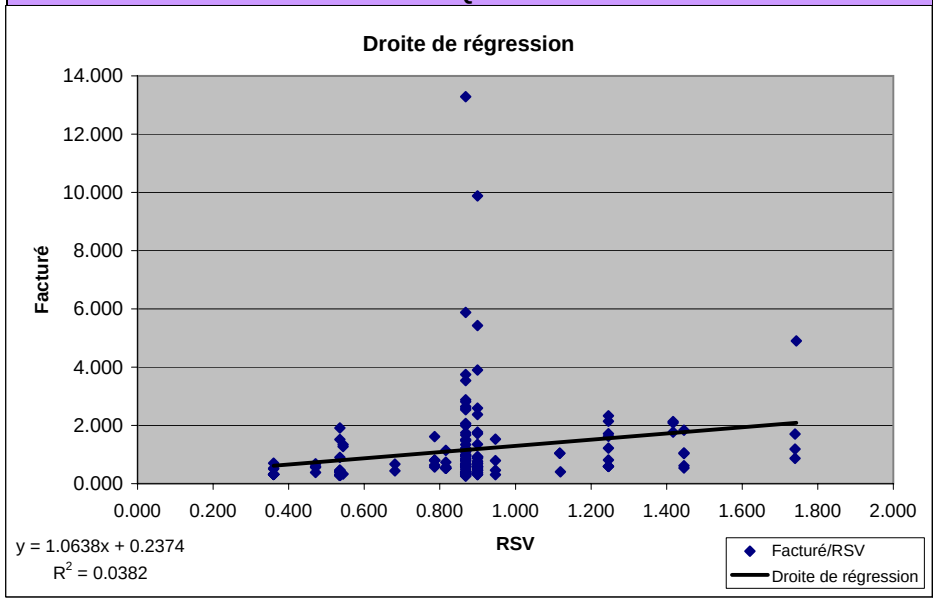
Fiche du test : Test_023 - CW

Identifiant du test : Test_023 - Date du test : 5/12/2007 - Modèle d'algorithme : MDT

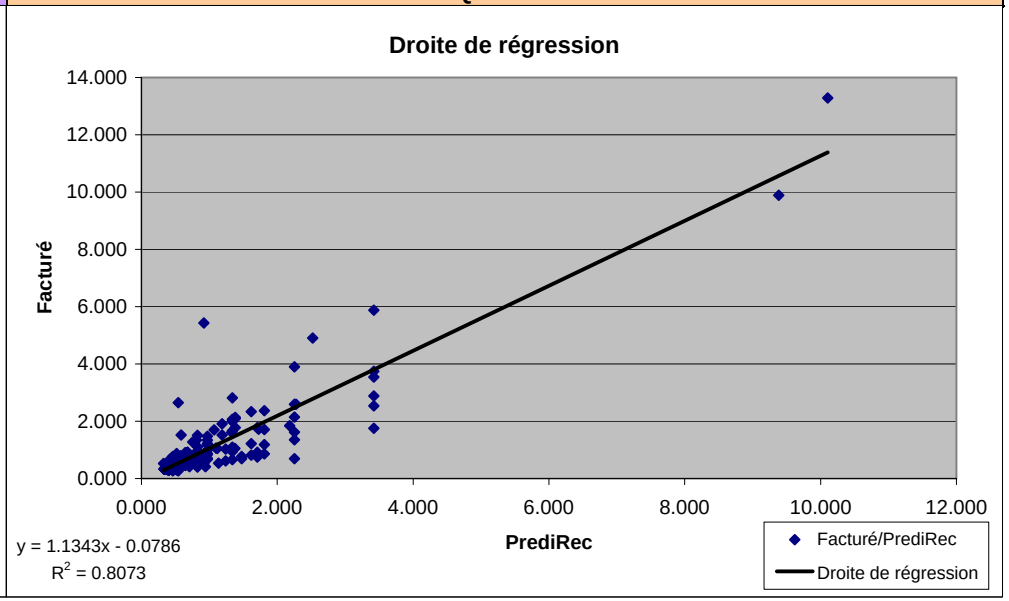
	Différence		Différence ABS	
	RSV	PrediRec	RSV	PrediRec
Minimum	- 12.421	- 4.513	0.004	-
Moyenne	- 0.293	- 0.068	0.678	0.376
Maximum	0.910	1.674	12.421	4.513
Total		185.34	138.54	174.48
			-33.79%	-6.23%

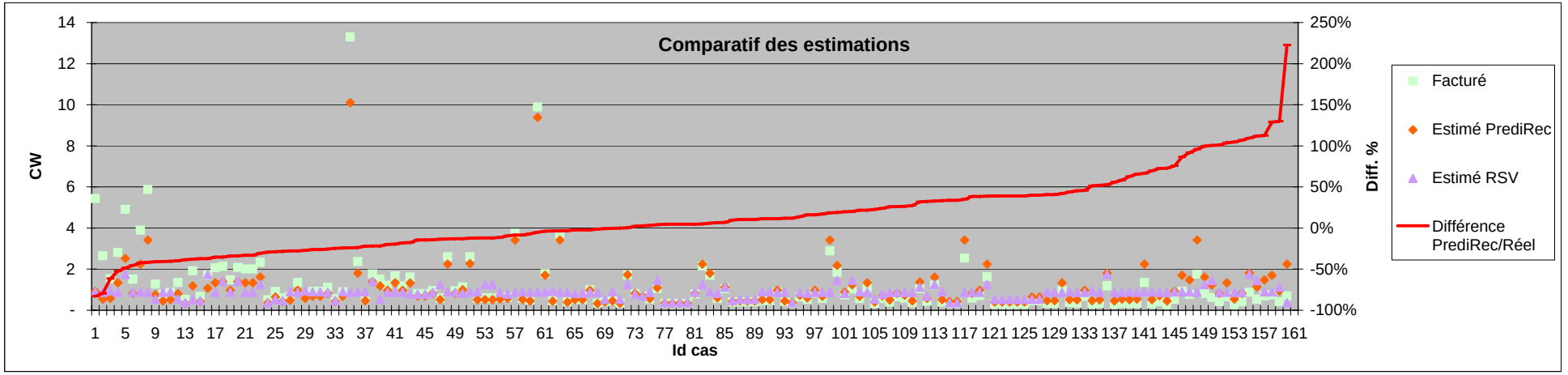
	Différence		Différence ABS		
	RSV	PrediRec	RSV	PrediRec	
Centile 10	- 1.219	- 0.637	0.038	0.023	
Centile 25	- 0.357	- 0.213	0.150	0.088	Quartile 1
Centile 50	0.076	0.037	0.350	0.184	Quartile 2
Centile 75	0.326	0.166	0.613	0.418	Quartile 3
Centile 90	0.540	0.534	1.219	0.781	

GRAPHIQUES RSV



GRAPHIQUES PREDIREC





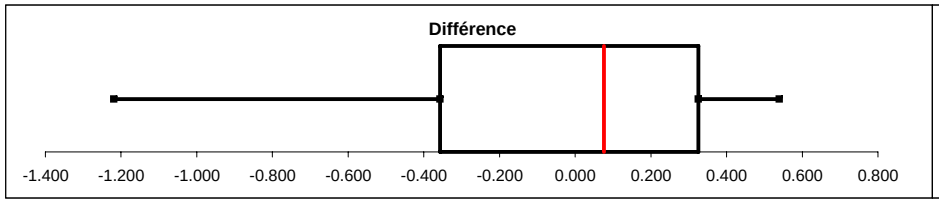
Id cas

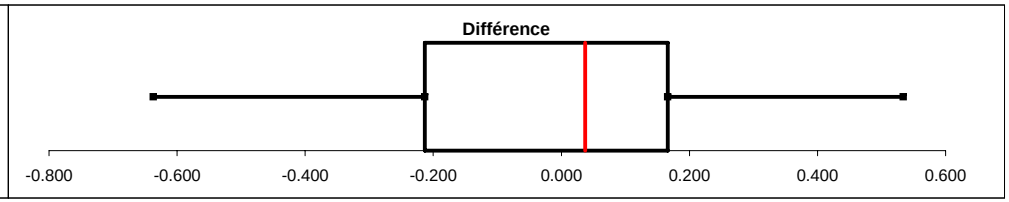
Fiche du test : Test_023 - CW

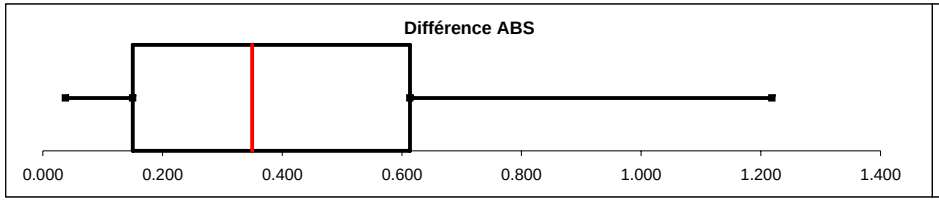
Identifiant du test : Test_023 - Date du test : 5/12/2007 - Modèle d'algorithme : MDT

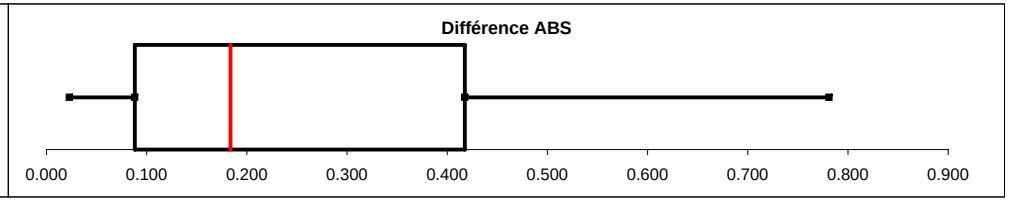
Données pour le "boxplot"

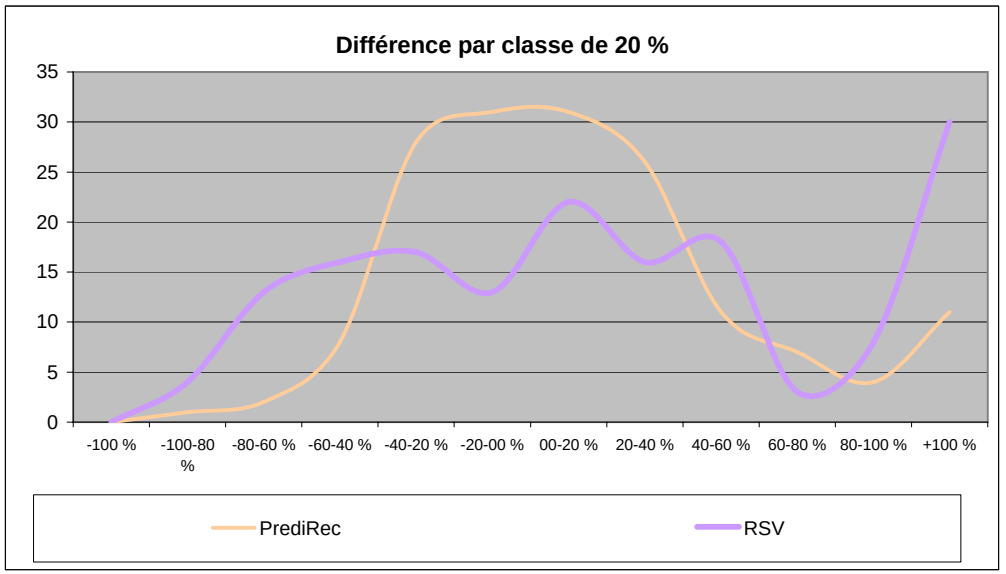
	C10	Q1	Q2	Q2	Q3	C90		Q1	Q1	Q3	Q3	Q1
Différence	1	1	0	2	1	1	**Base**	0	2	2	0	0
RSV	-1.219	-0.357	0.076	0.076	0.326	0.540	**RSV**	-0.357	-0.357	0.326	0.326	-0.357
PrediRec	-0.637	-0.213	0.037	0.037	0.166	0.534	**PrediRec**	-0.213	-0.213	0.166	0.166	-0.213
RSV ABS	0.038	0.150	0.350	0.350	0.613	1.219	**RSV ABS**	0.150	0.150	0.613	0.613	0.150
PrediRec ABS	0.023	0.088	0.184	0.184	0.418	0.781	**PrediRec ABS**	0.088	0.088	0.418	0.418	0.088











Projet SIMAV

Fiche de test

Facturation par APDRG :
Prédiction des recettes des cas non codés

Identifiant du test	Test_023
Date du test	05.12.2007
Modèle d'algorithme	MDT
Variables utilisées	Age Cas Classe Duree Sejour Duree Sejour Nette Groupe Classe Heures SI Sexe Tarif Type Admission Type Patient Type Taxe

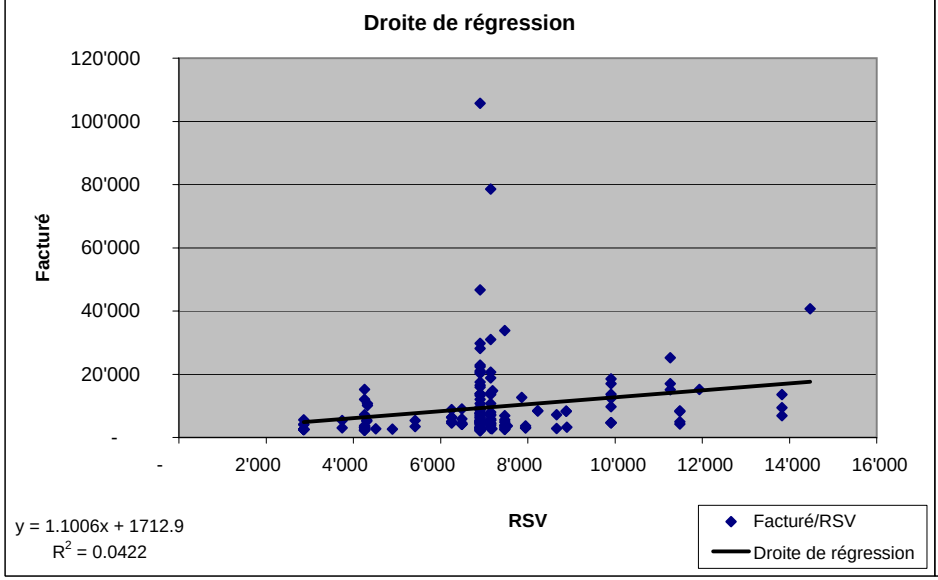
Fiche du test : Test_023 - Total facturé

Identifiant du test : Test_023 - Date du test : 5/12/2007 - Modèle d'algorithme : MDT

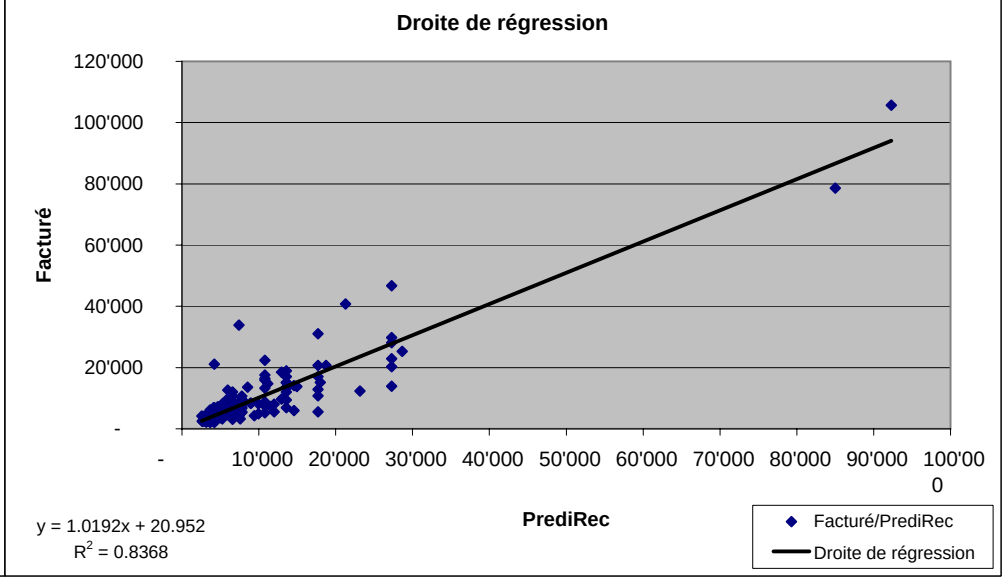
	Différence		Différence ABS	
	RSV	PrediRec	RSV	PrediRec
Minimum	- 98'799.69	- 26'435.32	50.69	12.17
Moyenne	- 2'414.75	- 197.59	5'452.77	2'830.19
Maximum	7'224.09	13'381.12	98'799.69	26'435.32
Total		1'502'618.65	1'116'259.01	1'471'004.33
			-34.61%	-2.15%

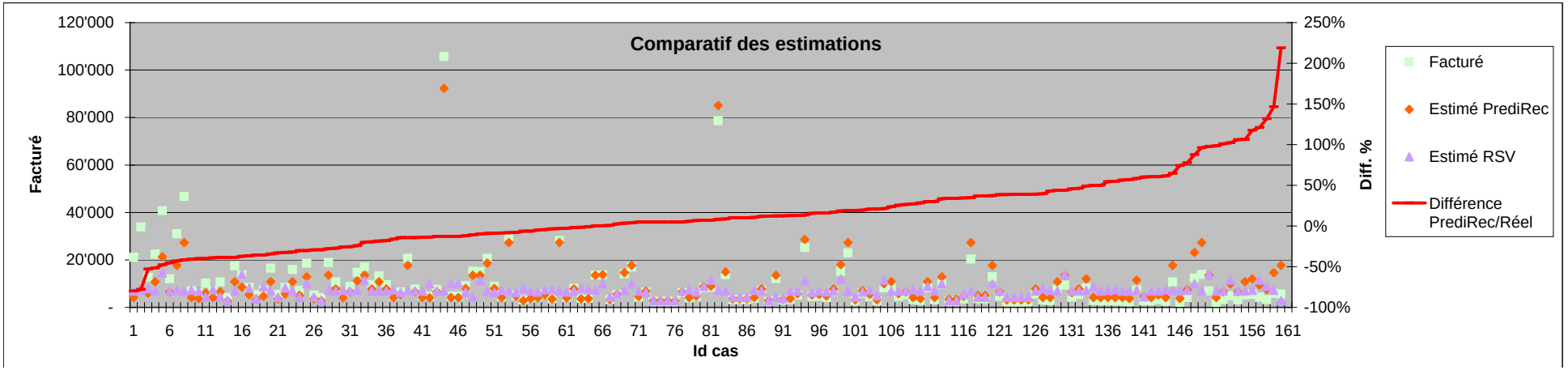
	Différence		Différence ABS		
	RSV	PrediRec	RSV	PrediRec	
Centile 10	- 11'020.92	- 4'088.60	303.27	224.17	
Centile 25	- 2'941.50	- 1'423.21	1'221.09	656.74	Quartile 1
Centile 50	588.40	329.41	2'677.06	1'478.32	Quartile 2
Centile 75	2'563.09	1'477.34	5'144.65	3'292.54	Quartile 3
Centile 90	4'415.18	4'162.55	11'020.92	6'499.34	

GRAPHIQUES RSV



GRAPHIQUES PREDIREC





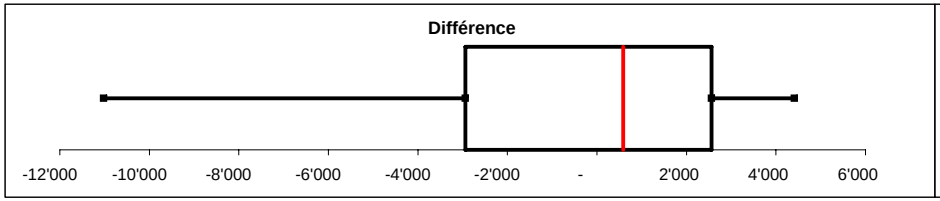
Id cas

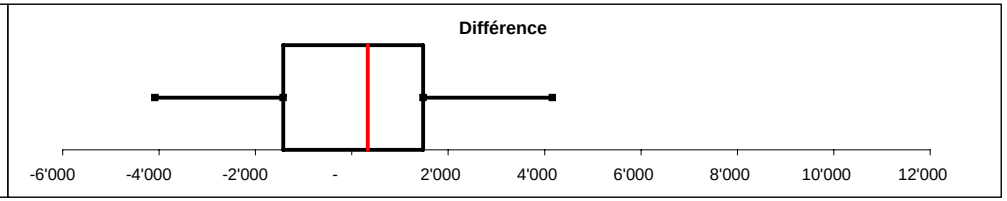
Fiche du test : Test_023 - Total facturé

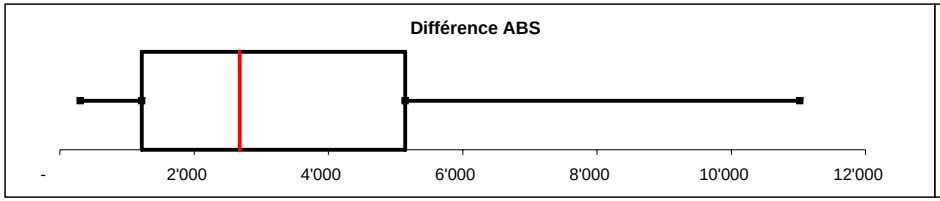
Identifiant du test : Test_023 - Date du test : 5/12/2007 - Modèle d'algorithme : MDT

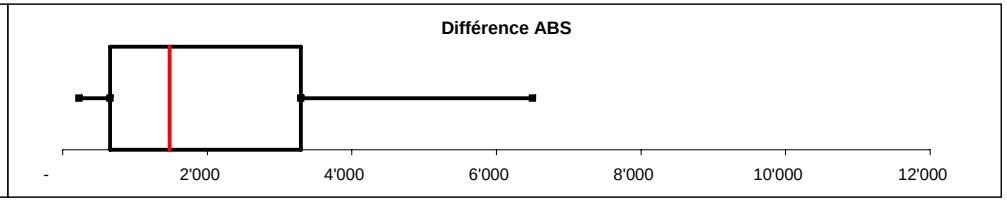
Données pour les "boxplot"

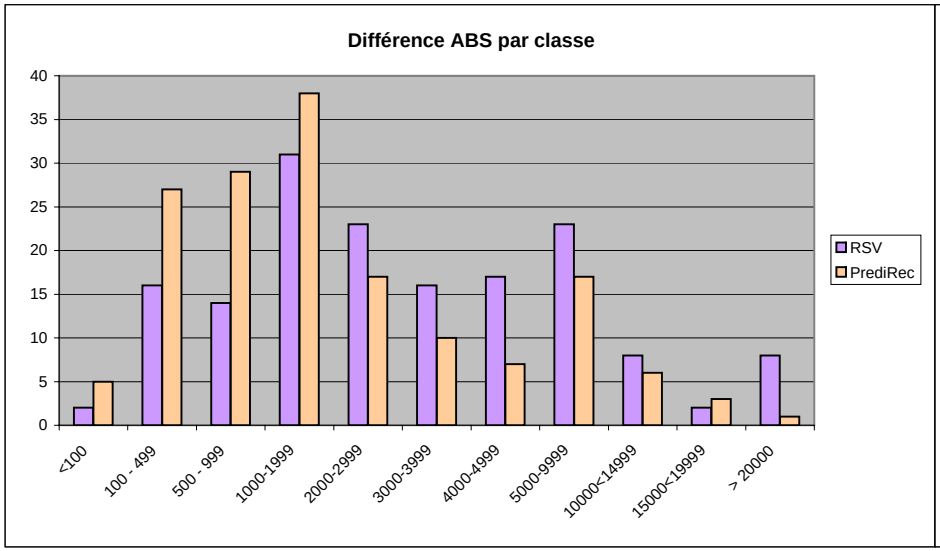
	C10	Q1	Q2	Q2	Q3	C90		Q1	Q1	Q3	Q3	Q1
Différence	1	1	0	2	1	1	**Base**	0	2	2	0	0
RSV	- 11'020.92	- 2'941.50	588.40	588.40	2'563.09	4'415.18	**RSV**	- 2'941.50	- 2'941.50	2'563.09	2'563.09	- 2'941.50
PrediRec	- 4'088.60	- 1'423.21	329.41	329.41	1'477.34	4'162.55	**PrediRec**	- 1'423.21	- 1'423.21	1'477.34	1'477.34	- 1'423.21
RSV ABS	303.27	1'221.09	2'677.06	2'677.06	5'144.65	11'020.92	**RSV ABS**	1'221.09	1'221.09	5'144.65	5'144.65	1'221.09
PrediRec ABS	224.17	656.74	1'478.32	1'478.32	3'292.54	6'499.34	**PrediRec ABS**	656.74	656.74	3'292.54	3'292.54	656.74

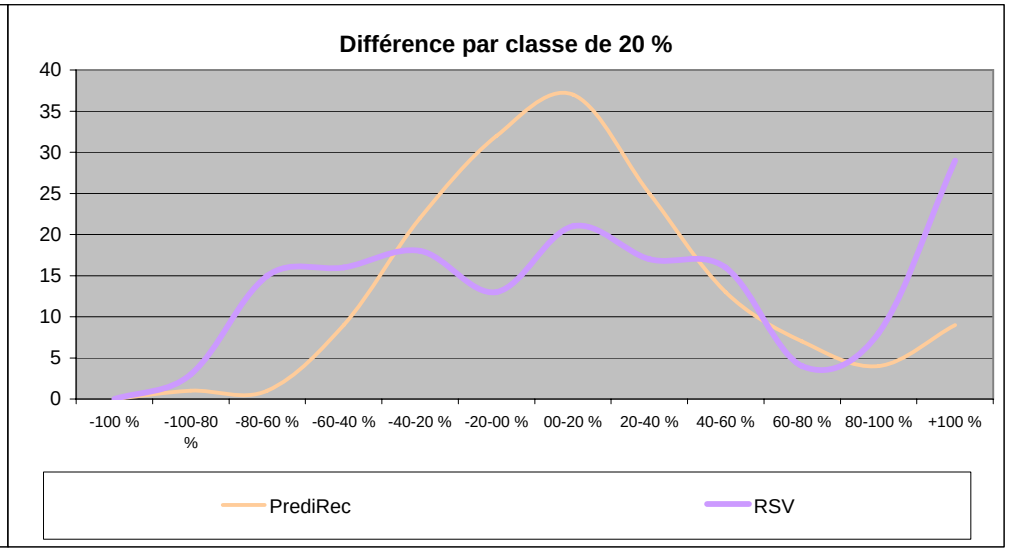












Synthèse des tests pour l'estimation du "Total Facture"

	Total Facture			Différence		Différence %		R2	
	Réel	RSV	PrediRec	RSV	PrediRec	RSV	PrediRec	RSV	PrediRec
PrediRec_Resultat_Test_025_04_MDT	1'502'618.65	1'116'259.01	1'440'071.73	- 386'359.64	- 62'546.92	-25.7%	-4.2%	R2 = 0.0422	R2 = 0.8611
PrediRec_Resultat_Test_024_04_MDT	1'502'618.65	1'116'259.01	1'447'251.70	- 386'359.64	- 55'366.95	-25.7%	-3.7%	R2 = 0.0422	R2 = 0.8572
PrediRec_Resultat_Test_026_04_MDT	1'502'618.65	1'116'259.01	1'462'230.01	- 386'359.64	- 40'388.64	-25.7%	-2.7%	R2 = 0.0422	R2 = 0.8437
PrediRec_Resultat_Test_023_MDT	1'502'618.65	1'116'259.01	1'471'004.33	- 386'359.64	- 31'614.32	-25.7%	-2.1%	R2 = 0.0422	R2 = 0.8368
PrediRec_Resultat_Test_010	1'339'587.65	934'676.78	1'279'834.76	- 404'910.87	- 59'752.89	-30.2%	-4.5%	R2 = 0.0339	R2 = 0.8023
PrediRec_Resultat_Test_004_04	1'502'618.65	1'116'259.01	1'459'385.37	- 386'359.64	- 43'233.28	-25.7%	-2.9%	R2 = 0.0422	R2 = 0.7873
PrediRec_Resultat_Test_006_04	1'502'618.65	1'116'259.01	1'476'701.49	- 386'359.64	- 25'917.16	-25.7%	-1.7%	R2 = 0.0422	R2 = 0.7867
PrediRec_Resultat_Test_005_04	1'502'618.65	1'116'259.01	1'480'275.44	- 386'359.64	- 22'343.21	-25.7%	-1.5%	R2 = 0.0422	R2 = 0.7832
PrediRec_Resultat_Test_023_MNN	1'502'618.65	1'116'259.01	1'410'801.71	- 386'359.64	- 91'816.94	-25.7%	-6.1%	R2 = 0.0422	R2 = 0.7827
PrediRec_Resultat_Test_006_02	1'502'618.65	1'116'259.01	1'455'972.29	- 386'359.64	- 46'646.36	-25.7%	-3.1%	R2 = 0.0422	R2 = 0.7818
PrediRec_Resultat_Test_004_02	1'502'618.65	1'116'259.01	1'464'440.49	- 386'359.64	- 38'178.16	-25.7%	-2.5%	R2 = 0.0422	R2 = 0.7816
PrediRec_Resultat_Test_006_01	1'502'618.65	1'116'259.01	1'452'870.32	- 386'359.64	- 49'748.33	-25.7%	-3.3%	R2 = 0.0422	R2 = 0.7776
PrediRec_Resultat_Test_005_02	1'502'618.65	1'116'259.01	1'460'617.46	- 386'359.64	- 42'001.19	-25.7%	-2.8%	R2 = 0.0422	R2 = 0.7748
PrediRec_Resultat_Test_005_01	1'502'618.65	1'116'259.01	1'460'617.46	- 386'359.64	- 42'001.19	-25.7%	-2.8%	R2 = 0.0422	R2 = 0.7748
PrediRec_Resultat_Test_025_04_MNN	1'502'618.65	1'116'259.01	1'409'340.15	- 386'359.64	- 93'278.50	-25.7%	-6.2%	R2 = 0.0422	R2 = 0.7676
PrediRec_Resultat_Test_026_04_MNN	1'502'618.65	1'116'259.01	1'393'608.04	- 386'359.64	- 109'010.61	-25.7%	-7.3%	R2 = 0.0422	R2 = 0.7668
PrediRec_Resultat_Test_011	1'339'587.65	934'676.78	1'270'713.46	- 404'910.87	- 68'874.19	-30.2%	-5.1%	R2 = 0.0339	R2 = 0.7624
PrediRec_Resultat_Test_030	1'502'618.65	1'116'259.01	1'423'907.41	- 386'359.64	- 78'711.24	-25.7%	-5.2%	R2 = 0.0422	R2 = 0.7567
PrediRec_Resultat_Test_024_04_MNN	1'502'618.65	1'116'259.01	1'435'893.87	- 386'359.64	- 66'724.78	-25.7%	-4.4%	R2 = 0.0422	R2 = 0.7515
PrediRec_Resultat_Test_009_03	1'502'618.65	1'116'259.01	1'446'141.53	- 386'359.64	- 56'477.12	-25.7%	-3.8%	R2 = 0.0422	R2 = 0.6517
PrediRec_Resultat_Test_009_02	1'502'618.65	1'116'259.01	1'446'141.53	- 386'359.64	- 56'477.12	-25.7%	-3.8%	R2 = 0.0422	R2 = 0.6517
PrediRec_Resultat_Test_003	1'502'618.65	1'116'259.01	1'450'214.40	- 386'359.64	- 52'404.25	-25.7%	-3.5%	R2 = 0.0422	R2 = 0.6505
PrediRec_Resultat_Test_007_02	1'502'618.65	1'116'259.01	1'428'611.30	- 386'359.64	- 74'007.35	-25.7%	-4.9%	R2 = 0.0422	R2 = 0.6491
PrediRec_Resultat_Test_009_01	1'502'618.65	1'116'259.01	1'453'854.38	- 386'359.64	- 48'764.27	-25.7%	-3.2%	R2 = 0.0422	R2 = 0.6483
PrediRec_Resultat_Test_004_03	1'502'618.65	1'116'259.01	1'437'464.81	- 386'359.64	- 65'153.84	-25.7%	-4.3%	R2 = 0.0422	R2 = 0.6469
PrediRec_Resultat_Test_008_02	1'502'618.65	1'116'259.01	1'446'857.21	- 386'359.64	- 55'761.44	-25.7%	-3.7%	R2 = 0.0422	R2 = 0.6468
PrediRec_Resultat_Test_004_01	1'502'618.65	1'116'259.01	1'442'515.71	- 386'359.64	- 60'102.94	-25.7%	-4.0%	R2 = 0.0422	R2 = 0.6462
PrediRec_Resultat_Test_002	1'502'618.65	1'116'259.01	1'442'515.71	- 386'359.64	- 60'102.94	-25.7%	-4.0%	R2 = 0.0422	R2 = 0.6462
PrediRec_Resultat_Test_001	1'502'618.65	1'116'259.01	1'442'515.71	- 386'359.64	- 60'102.94	-25.7%	-4.0%	R2 = 0.0422	R2 = 0.6462
PrediRec_Resultat_Test_007_01	1'502'618.65	1'116'259.01	1'433'244.11	- 386'359.64	- 69'374.54	-25.7%	-4.6%	R2 = 0.0422	R2 = 0.6452
PrediRec_Resultat_Test_008_01	1'502'618.65	1'116'259.01	1'459'130.37	- 386'359.64	- 43'488.28	-25.7%	-2.9%	R2 = 0.0422	R2 = 0.6434
PrediRec_Resultat_Test_005_03	1'502'618.65	1'116'259.01	1'433'969.61	- 386'359.64	- 68'649.04	-25.7%	-4.6%	R2 = 0.0422	R2 = 0.6429
PrediRec_Resultat_Test_006_03	1'502'618.65	1'116'259.01	1'441'088.49	- 386'359.64	- 61'530.16	-25.7%	-4.1%	R2 = 0.0422	R2 = 0.6421
PrediRec_Resultat_Test_007_03	1'502'618.65	1'116'259.01	1'425'020.31	- 386'359.64	- 77'598.34	-25.7%	-5.2%	R2 = 0.0422	R2 = 0.6398
PrediRec_Resultat_Test_008_03	1'502'618.65	1'116'259.01	1'451'049.68	- 386'359.64	- 51'568.97	-25.7%	-3.4%	R2 = 0.0422	R2 = 0.6394
PrediRec_Resultat_Test_014	1'339'587.65	934'676.78	1'195'170.39	- 404'910.87	- 144'417.26	-30.2%	-10.8%	R2 = 0.0339	R2 = 0.5975
PrediRec_Resultat_Test_013	1'339'587.65	934'676.78	1'089'081.24	- 404'910.87	- 250'506.41	-30.2%	-18.7%	R2 = 0.0339	R2 = 0.5927
PrediRec_Resultat_Test_012	1'339'587.65	934'676.78	1'089'081.24	- 404'910.87	- 250'506.41	-30.2%	-18.7%	R2 = 0.0339	R2 = 0.5927
PrediRec_Resultat_Test_015	1'339'587.65	934'676.78	1'115'654.78	- 404'910.87	- 223'932.87	-30.2%	-16.7%	R2 = 0.0339	R2 = 0.4156

Synthèse des tests pour l'estimation du "Cost-Weight pondéré"

	CW pondéré			Différence		Différence %		R2	
	Réel	RSV	PrediRec	RSV	PrediRec	RSV	PrediRec	RSV	PrediRec
PrediRec_Resultat_Test_011	164.95	115.945	169.011	-49.003	4.063	-29.7%	2.5%	R2 = 0.0301	R2 = 0.8503
PrediRec_Resultat_Test_010	164.95	115.945	156.936	-49.003	-8.012	-29.7%	-4.9%	R2 = 0.0301	R2 = 0.8084
PrediRec_Resultat_Test_023_MDT	185.34	138.535	174.481	-46.809	-10.863	-25.3%	-5.9%	R2 = 0.0382	R2 = 0.8073
PrediRec_Resultat_Test_005_04	185.34	138.535	167.182	-46.809	-18.162	-25.3%	-9.8%	R2 = 0.0382	R2 = 0.7815
PrediRec_Resultat_Test_025_04_MDT	185.34	138.535	169.598	-46.809	-15.746	-25.3%	-8.5%	R2 = 0.0382	R2 = 0.7787
PrediRec_Resultat_Test_008_01	185.34	138.535	167.796	-46.809	-17.548	-25.3%	-9.5%	R2 = 0.0382	R2 = 0.7755
PrediRec_Resultat_Test_005_03	185.34	138.535	163.026	-46.809	-22.318	-25.3%	-12.0%	R2 = 0.0382	R2 = 0.7751
PrediRec_Resultat_Test_024_04_MDT	185.34	138.535	169.971	-46.809	-15.373	-25.3%	-8.3%	R2 = 0.0382	R2 = 0.7733
PrediRec_Resultat_Test_008_02	185.34	138.535	164.462	-46.809	-20.882	-25.3%	-11.3%	R2 = 0.0382	R2 = 0.7691
PrediRec_Resultat_Test_008_03	185.34	138.535	168.232	-46.809	-17.112	-25.3%	-9.2%	R2 = 0.0382	R2 = 0.7686
PrediRec_Resultat_Test_005_02	185.34	138.535	164.432	-46.809	-20.912	-25.3%	-11.3%	R2 = 0.0382	R2 = 0.7655
PrediRec_Resultat_Test_005_01	185.34	138.535	164.432	-46.809	-20.912	-25.3%	-11.3%	R2 = 0.0382	R2 = 0.7655
PrediRec_Resultat_Test_026_04_MDT	185.34	138.535	171.326	-46.809	-14.018	-25.3%	-7.6%	R2 = 0.0382	R2 = 0.7604
PrediRec_Resultat_Test_025_04_MNN	185.34	138.535	164.759	-46.809	-20.585	-25.3%	-11.1%	R2 = 0.0382	R2 = 0.7592
PrediRec_Resultat_Test_026_04_MNN	185.34	138.535	166.271	-46.809	-19.073	-25.3%	-10.3%	R2 = 0.0382	R2 = 0.7545
PrediRec_Resultat_Test_023_MNN	185.34	138.535	168.810	-46.809	-16.534	-25.3%	-8.9%	R2 = 0.0382	R2 = 0.754
PrediRec_Resultat_Test_007_01	185.34	138.535	164.664	-46.809	-20.680	-25.3%	-11.2%	R2 = 0.0382	R2 = 0.7529
PrediRec_Resultat_Test_007_03	185.34	138.535	164.633	-46.809	-20.711	-25.3%	-11.2%	R2 = 0.0382	R2 = 0.7527
PrediRec_Resultat_Test_006_04	185.34	138.535	166.201	-46.809	-19.143	-25.3%	-10.3%	R2 = 0.0382	R2 = 0.7483
PrediRec_Resultat_Test_006_02	185.34	138.535	163.324	-46.809	-22.020	-25.3%	-11.9%	R2 = 0.0382	R2 = 0.746
PrediRec_Resultat_Test_009_03	185.34	138.535	166.589	-46.809	-18.755	-25.3%	-10.1%	R2 = 0.0382	R2 = 0.7454
PrediRec_Resultat_Test_009_02	185.34	138.535	166.589	-46.809	-18.755	-25.3%	-10.1%	R2 = 0.0382	R2 = 0.7454
PrediRec_Resultat_Test_004_01	185.34	138.535	164.636	-46.809	-20.708	-25.3%	-11.2%	R2 = 0.0382	R2 = 0.7436
PrediRec_Resultat_Test_002	185.34	138.535	164.636	-46.809	-20.708	-25.3%	-11.2%	R2 = 0.0382	R2 = 0.7436
PrediRec_Resultat_Test_001	185.34	138.535	164.636	-46.809	-20.708	-25.3%	-11.2%	R2 = 0.0382	R2 = 0.7436
PrediRec_Resultat_Test_030	185.34	138.535	167.432	-46.809	-17.912	-25.3%	-9.7%	R2 = 0.0382	R2 = 0.7427
PrediRec_Resultat_Test_006_01	185.34	138.535	163.916	-46.809	-21.428	-25.3%	-11.6%	R2 = 0.0382	R2 = 0.7425
PrediRec_Resultat_Test_006_03	185.34	138.535	164.055	-46.809	-21.289	-25.3%	-11.5%	R2 = 0.0382	R2 = 0.7421
PrediRec_Resultat_Test_009_01	185.34	138.535	168.440	-46.809	-16.904	-25.3%	-9.1%	R2 = 0.0382	R2 = 0.7391
PrediRec_Resultat_Test_003	185.34	138.535	163.746	-46.809	-21.598	-25.3%	-11.7%	R2 = 0.0382	R2 = 0.7354
PrediRec_Resultat_Test_007_02	185.34	138.535	162.821	-46.809	-22.523	-25.3%	-12.2%	R2 = 0.0382	R2 = 0.7297
PrediRec_Resultat_Test_024_04_MNN	185.34	138.535	169.224	-46.809	-16.120	-25.3%	-8.7%	R2 = 0.0382	R2 = 0.7281
PrediRec_Resultat_Test_004_02	185.34	138.535	165.065	-46.809	-20.279	-25.3%	-10.9%	R2 = 0.0382	R2 = 0.7274
PrediRec_Resultat_Test_004_04	185.34	138.535	165.987	-46.809	-19.357	-25.3%	-10.4%	R2 = 0.0382	R2 = 0.7253
PrediRec_Resultat_Test_004_03	185.34	138.535	160.979	-46.809	-24.365	-25.3%	-13.1%	R2 = 0.0382	R2 = 0.6646
PrediRec_Resultat_Test_014	164.95	115.945	143.378	-49.003	-21.570	-29.7%	-13.1%	R2 = 0.0301	R2 = 0.6011
PrediRec_Resultat_Test_013	164.95	115.945	136.252	-49.003	-28.696	-29.7%	-17.4%	R2 = 0.0301	R2 = 0.5326
PrediRec_Resultat_Test_012	164.95	115.945	136.252	-49.003	-28.696	-29.7%	-17.4%	R2 = 0.0301	R2 = 0.5326
PrediRec_Resultat_Test_015	164.95	115.945	134.425	-49.003	-30.523	-29.7%	-18.5%	R2 = 0.0301	R2 = 0.4889